



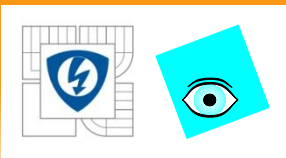
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Počítačová analýza vícerozměrných dat v oborech přírodních, technických a společenských věd

Prof. RNDr. Milan Meloun, DrSc. (Univerzita Pardubice, Pardubice)

20.-24. června 2011

Tato prezentace je spolufinancována Evropským sociálním fondem a státním rozpočtem České republiky.

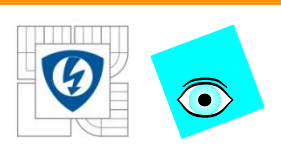


4.9 ANALÝZA SHLUKŮ

25.2.2010

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

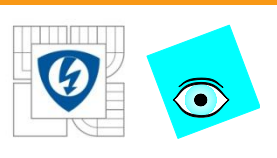




Postup analýzy shluků

Poskytuje empirické a objektivní metody ke klasifikaci objektů

- 1. krok: Cíle analýzy shluků**
- 2. krok: Formulace úlohy analýzy shluků**
- 3. krok: Předpoklady analýzy shluků**
- 4. krok: Výstavba dendrogramu shluků**
- 5. krok: Interpretace shluků**
- 6. krok: Validace a profilování shluků**



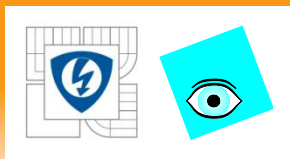
1. krok: Cíle analýzy shluků

Rozdělení objektů do shluků dle podobnosti objektů
a dle specifikovaných vlastností - proměnných.

Popis systematiky (taxonomie): empirická klasifikace. Shluky objektů jsou porovnány s jejich teoretickou typologií.

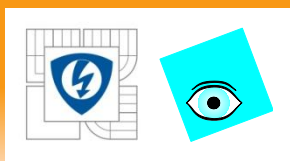
Zjednodušení dat: zjednodušený pohled na soubor objektů. Na oddělené shluky objektů se hledí dle jejich vlastností.

Identifikace vztahu: dle struktury shluků je snadnější odhalit vztahy mezi objekty. Shluky mohou být předmětem dalšího kvalitativního uvažování.



Úloha 1. Klasifikace polétavých mšic (Kompendium B404)

Jeffers (1967)25 studoval 40 jedinců polétavých mšic (*Alate adelges*): 19 ukazatelů k rozlišení druhů, 14 znaků délky a šířky, 4 znaky se týkají počtu a 1 binární vyjadřuje přítomnost či absenci: **x1** délka těla, **x2** šířka těla, **x3** délka předního křídla, **x4** délka zadního křídla, **x5** počet průduchů, **x6** délka tykadla I, **x7** délka tykadla II, **x8** délka tykadla III, **x9** délka tykadla IV, **x10** délka tykadla V, **x11** počet tykadlových ostnů, **x12** délka posledního článku nohy, **x13** délka holeně, tibia, **x14** délka stehna, **x15** délka sosáku, **x16** délka kladélka, **x17** počet kladélkových trnů, **x18** řitní otvor, **x19** počet háčků zadních křídel



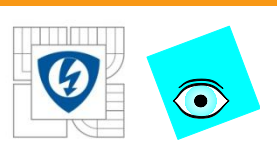
Úloha 1. Klasifikace polétavých mšic (Kompendium B404)

x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15	x16	x17	x18	x19
21.2	11	7.5	4.8	5	2	2	2.8	2.8	3.3	3	4.4	4.5	3.6	7	4	8	0	3
20.2	10	7.5	5	5	2.3	2.1	3	3	3.2	5	4.2	4.5	3.5	7.6	4.2	8	0	3
20.2	10	7	4.6	5	1.9	2.1	3	2.5	3.3	1	4.2	4.4	3.3	7	4	6	0	3
22.5	8.8	7.4	4.7	5	2.4	2.1	3	2.7	3.5	5	4.2	4.4	3.6	6.8	4.1	6	0	3
20.6	11	8	4.8	5	2.4	2	2.9	2.7	3	4	4.2	4.7	3.5	6.7	4	6	0	3
19.1	9.2	7	4.5	5	1.8	1.9	2.8	3	3.2	5	4.1	4.3	3.3	5.7	3.8	8	0	3.5
20.8	11.4	7.7	4.9	5	2.5	2.1	3.1	3.1	3.2	4	4.2	4.7	3.6	6.6	4	8	0	3
15.5	8.2	6.3	4.9	5	2	2	2.9	2.4	3	3	3.7	3.8	2.9	6.7	3.5	6	0	3.5
16.7	8.8	6.4	4.5	5	2.1	1.9	2.8	2.7	3.1	3	3.7	3.8	2.8	6.1	3.7	8	0	3
19.7	9.9	8.2	4.7	5	2.2	2	3	3	3.1	0	4.1	4.3	3.3	6	3.8	8	0	3
10.6	5.2	3.9	2.3	4	1.2	1	2	2	2.2	6	2.5	2.5	2	4.5	2.7	4	1	2
9.2	4.5	3.7	2.2	4	1.3	1.2	2	1.6	2.1	5	2.4	2.3	1.8	4.1	2.4	4	1	2
9.6	4.5	3.6	2.3	4	1.3	1	1.9	1.7	2.2	4	2.4	2.3	1.7	4	2.3	4	1	2
8.5	4	3.8	2.2	4	1.3	1.1	1.9	2	2.1	5	2.4	2.4	1.9	4.4	2.3	4	1	2
11	4.7	4.2	2.3	4	1.2	1	1.9	2	2.2	4	2.5	2.5	2	4.5	2.6	4	1	2
18.1	8.2	5.9	3.5	5	1.9	1.9	1.9	2.7	2.8	4	3.5	3.8	2.9	6	4.5	9	1	2
17.6	8.3	6	3.8	5	2	1.9	2	2.2	2.9	3	3.5	3.6	2.8	5.7	4.3	10	1	2
19.2	6.6	6.2	3.4	5	2	1.8	2.2	2.3	2.8	4	3.5	3.4	2.5	5.3	3.8	10	1	2
15.4	7.6	7.1	3.4	5	2	1.9	2.5	2.5	2.9	4	3.3	3.6	2.7	6	4.2	8	1	3
15.1	7.3	6.2	3.8	5	2	1.8	2.1	2.4	2.5	4	3.7	3.7	2.8	6.4	4.3	10	1	2.5
16.1	7.9	5.8	3.7	5	2.1	1.9	2.3	2.6	2.9	5	3.6	3.6	2.7	6	4.5	0	1	2
19.1	8.8	6.4	3.9	5	2.2	2	2.3	2.4	2.9	4	3.8	4	3	6.5	4.5	0	1	2.5
15.3	6.4	5.3	3.3	5	1.7	1.6	2	2.2	2.5	5	3.4	3.4	2.6	5.4	4	0	1	2
14.8	8.1	6.2	3.7	5	2.2	2	2.2	2.4	3.2	5	3.5	3.7	2.7	6	4.1	0	1	2
16.2	7.7	6.9	3.7	5	2	1.8	2.3	2.4	2.8	4	3.8	3.7	2.7	5.7	4.2	0	1	2.5
13.4	6.9	5.7	3.4	5	2	1.8	2.8	2	2.6	4	3.6	3.6	2.6	5.5	3.9	0	1	2
12.9	5.8	4.8	2.6	5	1.6	1.5	1.9	2.1	2.6	5	2.8	3	2.2	5.1	3.6	9	1	3
12	6.5	5.3	3.2	5	1.9	1.9	2.3	2.5	3	5	3.3	3.5	2.6	5.4	4.3	8	1	2
14.1	7	5.5	3.6	5	2.2	2	2.3	2.5	3.1	5	3.6	3.7	2.8	5.8	4.1	0	1	2
16.7	7.2	5.7	3.5	5	1.9	1.9	2.5	2.3	2.8	5	3.4	3.6	2.7	6	4	0	1	2.5
14.1	5.4	5	3	5	1.7	1.6	1.8	2.5	2.4	5	2.7	2.9	2.2	5.3	3.6	8	1	2
10	6	4.2	2.5	5	1.6	1.4	1.4	2	2.7	6	2.8	2.5	1.8	4.8	3.4	8	1	2
11.4	4.5	4.4	2.7	5	1.8	1.5	1.9	1.7	2.5	5	2.7	2.5	1.9	4.7	3.7	8	1	2
12.5	5.5	4.7	2.3	5	1.8	1.4	1.8	2.2	2.4	4	2.8	2.6	2	5.1	3.7	8	0	2
13	5.3	4.7	2.3	5	1.6	1.4	1.8	1.8	2.5	4	2.7	2.7	2.1	5	3.6	8	1	2
12.4	5.2	4.4	2.6	5	1.6	1.4	1.8	2.2	2.2	5	2.7	2.5	2	5	3.2	6	1	2
12	5.4	4.9	3	5	1.7	1.5	1.7	1.9	2.4	5	2.7	2.7	2	4.2	3.7	6	1	2
10.7	5.6	4.5	2.8	5	1.8	1.4	1.8	2.2	2.4	4	2.7	2.6	2	5	3.5	8	1	2
11.1	5.5	4.3	2.6	5	1.7	1.5	1.8	1.9	2.4	5	2.6	2.5	1.9	4.6	3.4	8	1	2
12.8	5.7	4.8	2.8	5	1.6	1.4	1.7	1.9	2.3	5	2.3	2.5	1.9	5	3.1	8	1	2

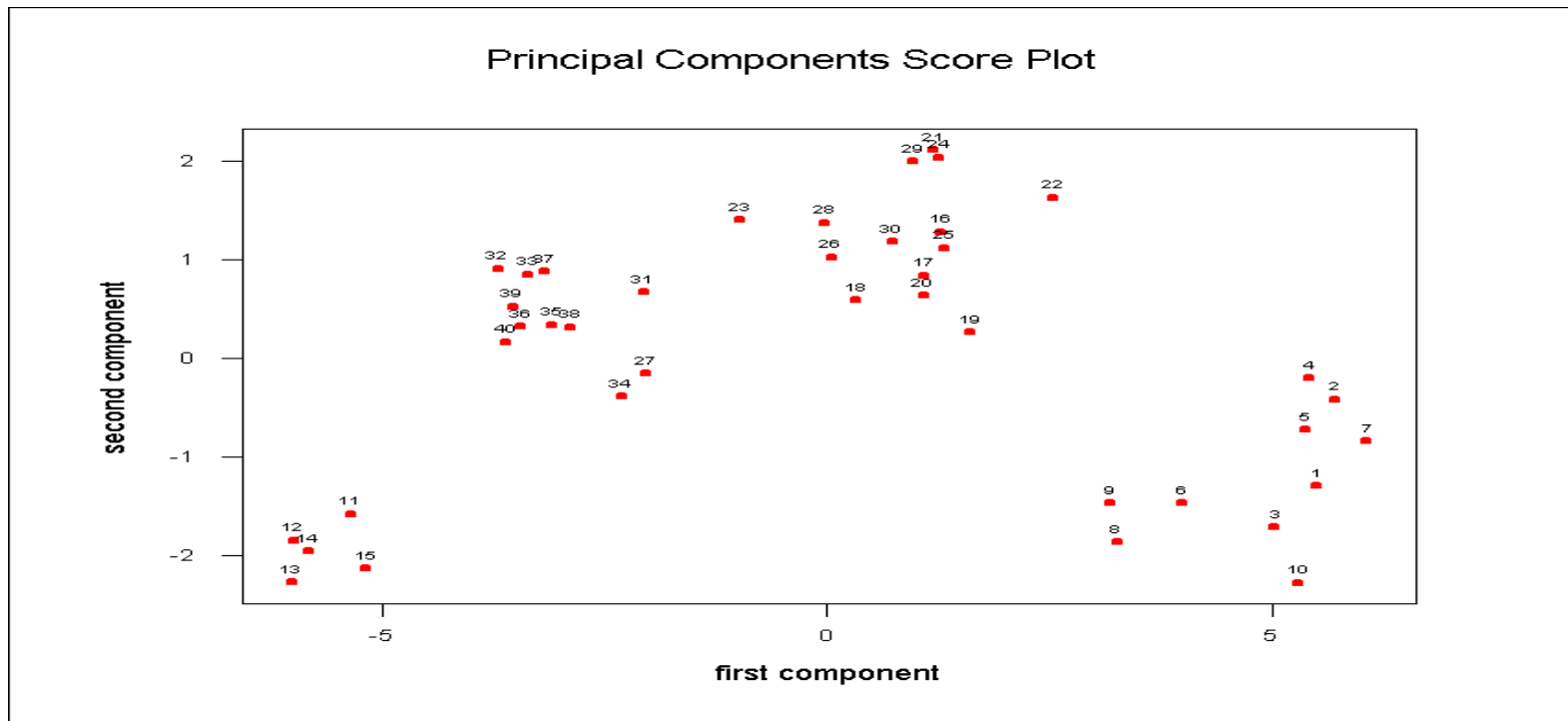
25.2.2010

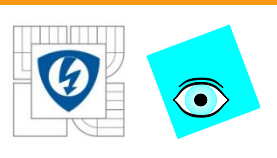
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ



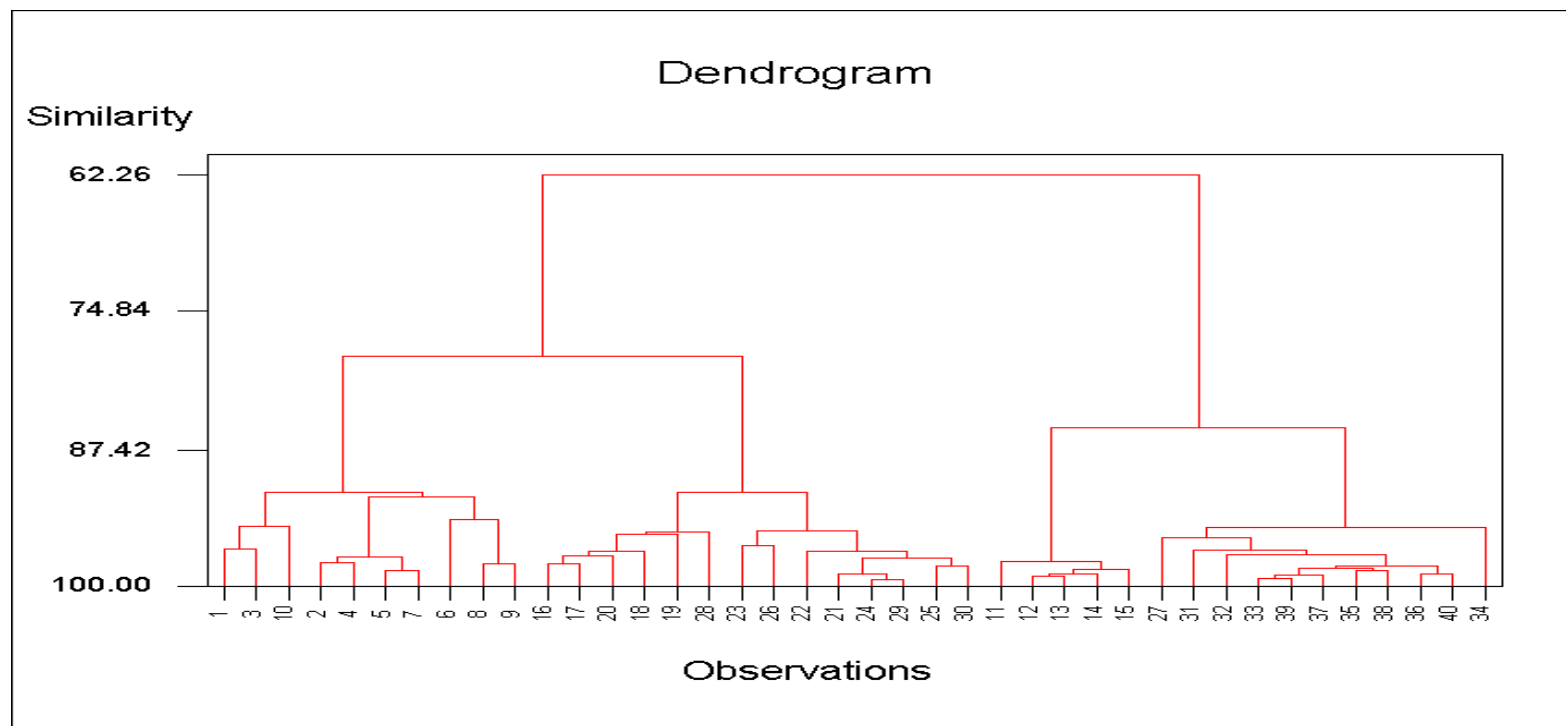


Úloha 1. Klasifikace polétavých mšic (Kompendium B404)





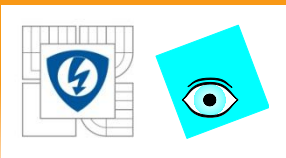
Úloha 1. Klasifikace polétavých mšic (Kompendium B404)



25.2.2010

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ



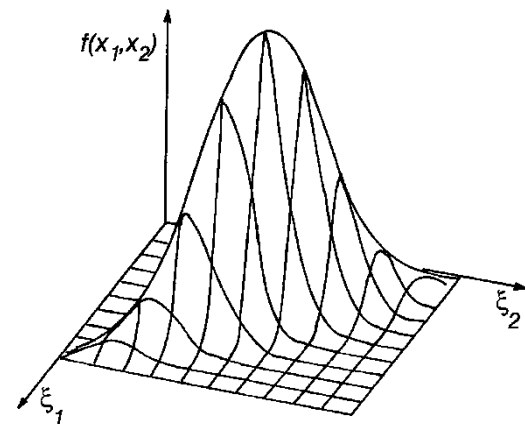
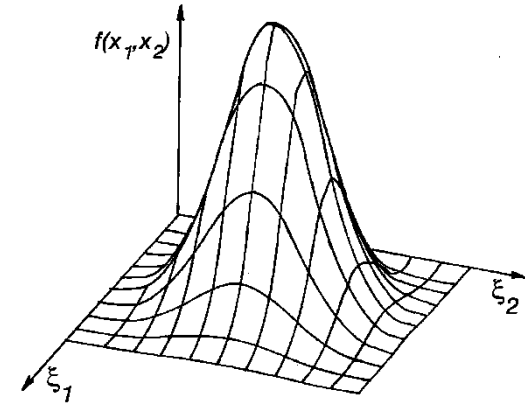


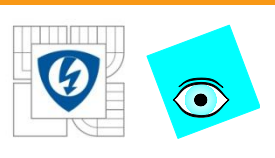
Volba shlukovacích proměnných či znaků (shluky vystihují strukturu dat dle nadefinovaných proměnných, znaků)

dle teoretických a praktických hledisek:

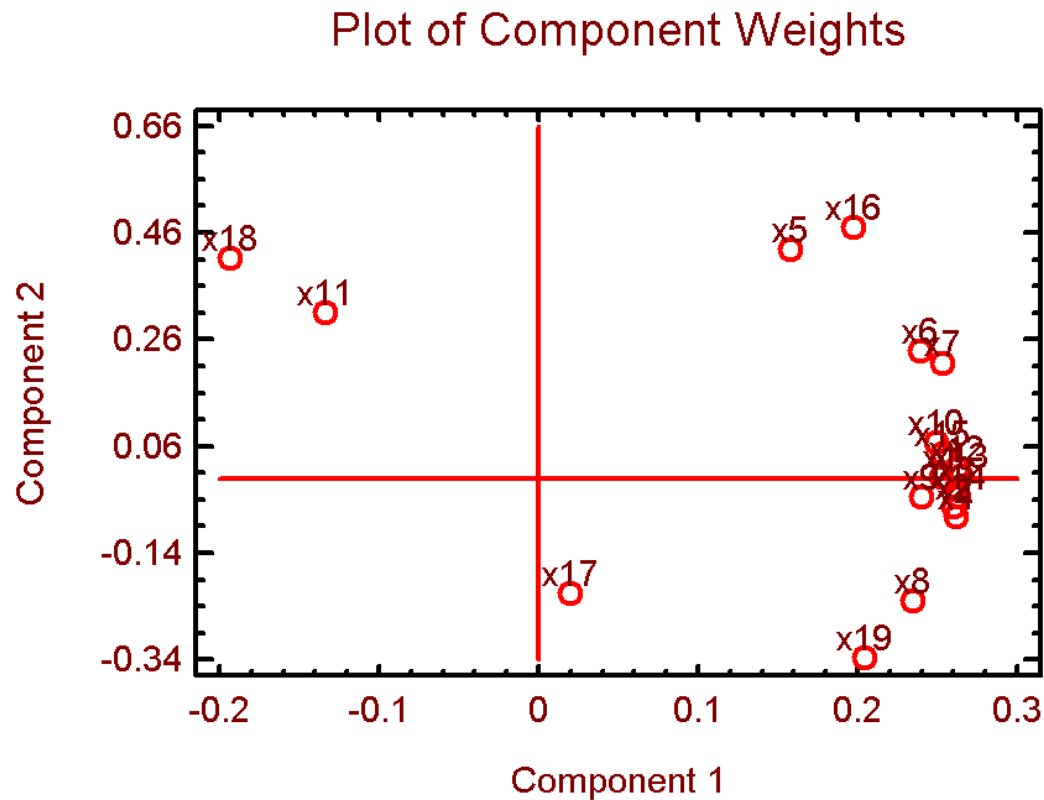
- 1) Proměnné charakterizují objekty shlukované.
- 2) Proměnné se vztahují k objektům shlukové analýzy.
- 3) Analýza nerozlišuje významné a nevýznamné proměnné.
- 4) Odlišení shluků za použití všech navržených proměnných.
- 5) Nevýznamné proměnné zvýší šanci odlehlých bodů.
- 6) Na volbě proměnných závisí nalezení správných shluků.
- 7) Pouze proměnné, které dostatečně rozlišují mezi objekty.

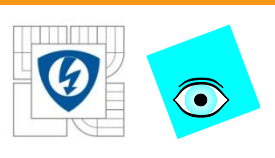
$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_j^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} = \begin{bmatrix} x_{1,1} & \dots & x_{1,i} & \dots & x_{1,n} \\ \vdots & & \vdots & & \vdots \\ x_{j,1} & \dots & x_{j,i} & \dots & x_{j,n} \\ \vdots & & \vdots & & \vdots \\ x_{n,1} & \dots & x_{n,i} & \dots & x_{n,n} \end{bmatrix}$$



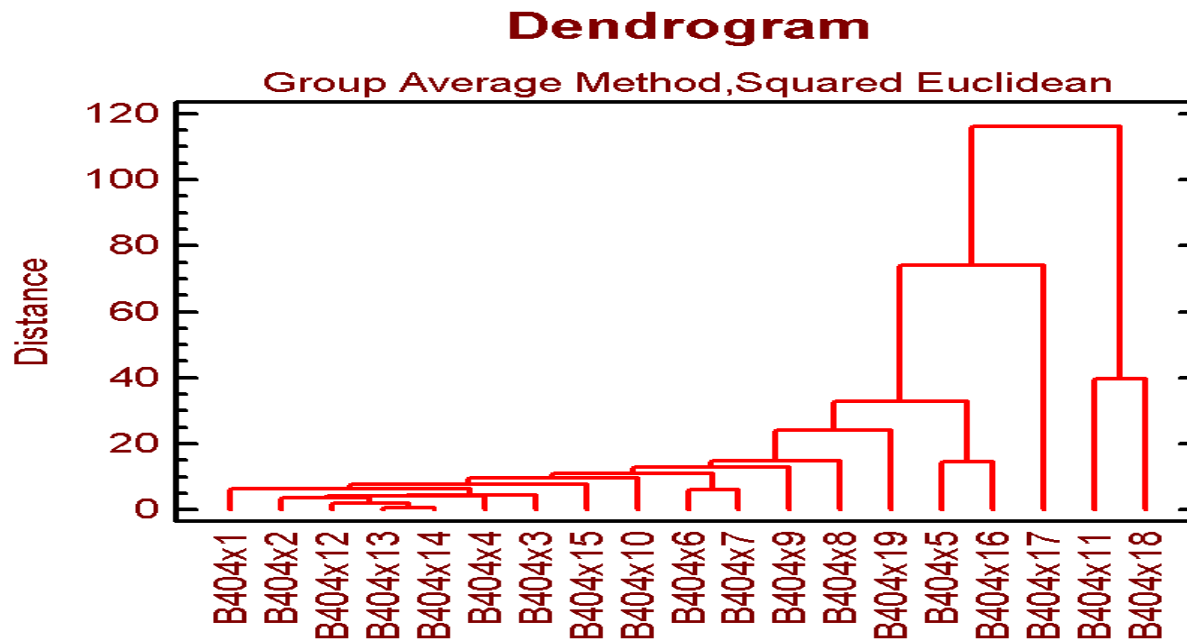


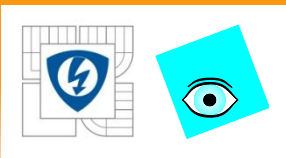
Plot of Component Weights





Dendrogram





Postup analýzy shluků

Poskytuje empirické a objektivní metody ke klasifikaci objektů

1. krok: Cíle analýzy shluků

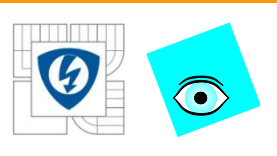
2. krok: Formulace úlohy analýzy shluků

3. krok: Předpoklady analýzy shluků

4. krok: Výstavba dendrogramu shluků

5. krok: Interpretace shluků

6. krok: Validace a profilování shluků



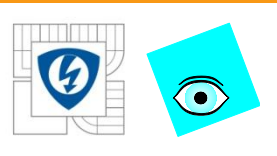
2. krok: Formulace úlohy analýzy shluků

Vyhodnotí všechny možné kombinace shluků, u 25 objektů a 5 shluků existuje 2.4×10^{15} možných shluků. Uživatel musí určit jediné správné řešení. Návrh modelu shluků a použité techniky má větší důležitost než u ostatních vícerozměrných technik.

2.1 Odhalení odlehlých objektů, outlierů

Outliery představují

- (1) odchýlené objekty, které nejsou představiteli populace,
- (2) chybný výběr objektu z dané populace.



Outliery

Outliery zbortí

- (1) strukturu dat,
- (2) nalezené shluky nebudou představovat skutečnou strukturu objektů dané populace.

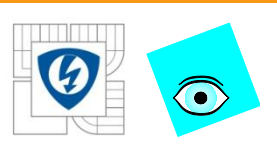
Nalezení outlierů

profilovým diagramem proměnných.

Outliery z dat odstranit

někdy se zbortí aktuální struktura objektů.





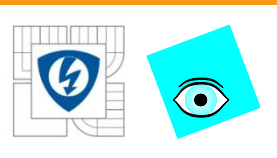
2.2 Standardizace dat

Aplikuje se před analýzou, míry vzdálenosti jsou citlivé na rozličné stupnice nebo na lišící se numerické velikosti proměnných.

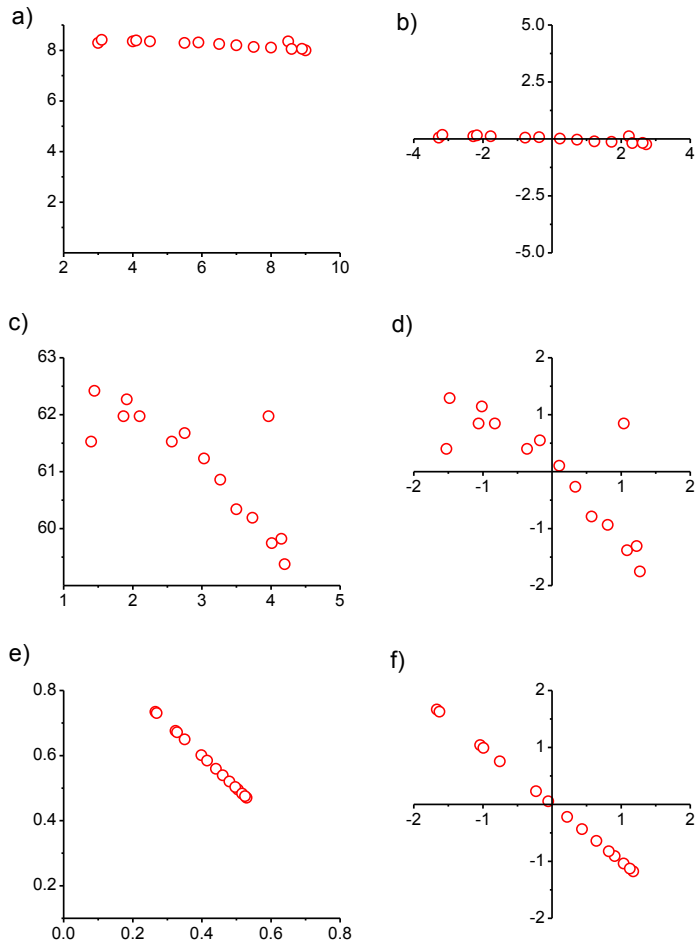
Pravidlo: proměnné s větší proměnlivostí (směrodatnou odchylkou) mají větší vliv na míru podobnosti.

Standardizování proměnných

Standardizace je transformace proměnné do svého Z-skóre: (odečtením sloupcového průměru od každé hodnoty ve sloupci a výsledek se podělí sloupcovou směrodatnou odchylkou): **Průměr standardizovaných dat je 0 se směrodatnou odchylkou 1.**

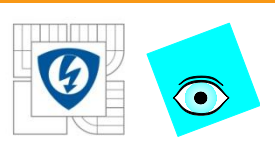


Efekt škálovacích technik



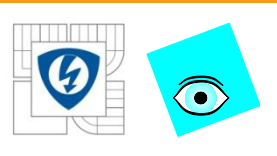
Efekt škálovacích technik:

- (a) Originální data,
- (b) sloupcové centrování,
- (c) sloupcové standardizování,
- (d) autoškálování,
- (e) profily,
- (f) autoškálované profily



Efekt škálovacích technik

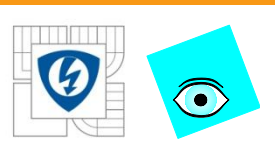
- 1) Sloupcové centrování dle $y_{ij} = x_{ij} - \bar{x}_j$.
- 2) Sloupcová standardizace dle $y_{ij} = x_{ij}/s_j$.
- 3) Autoškálování je tzv. studentizace dle $y_{ij} = (x_{ij} - \bar{x}_j)/s_j$ která je analogická Z-transformaci pro velké výběry $y_{ij} = (x_{ij} - \mu_j)/\sigma_j$.
- 4) Škálování sloupcovým rozsahem $y_{ij} = \frac{x_{ij} - \min_j x_{ij}}{\max_j x_{ij} - \min_j x_{ij}}$.
- 5) Řádkové centrování dle $y_{ij} = x_{ij} - \bar{x}_i$.
- 6) Řádková standardizace dle $y_{ij} = x_{ij}/s_i$.
- 7) Celkové centrování $y_{ij} = x_{ij} - \bar{x}$, kde $\bar{x}$ je celkový průměr.
- 8) Celková standardizace dle $y_{ij} = x_{ij}/s$, kde s je směrodatná odchylka
- 9) Řádkové profily dle $y_{ij} = x_{ij}/(\bar{x}_i m)$.
- 10) Sloupcové profily dle $y_{ij} = x_{ij}/(\bar{x}_j n)$.



Vlastnosti

Vlastnosti:

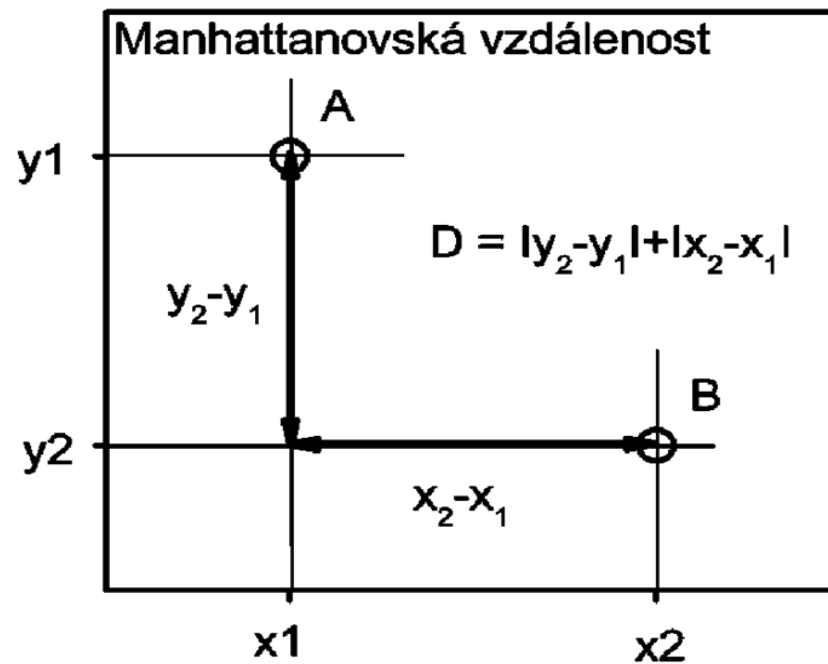
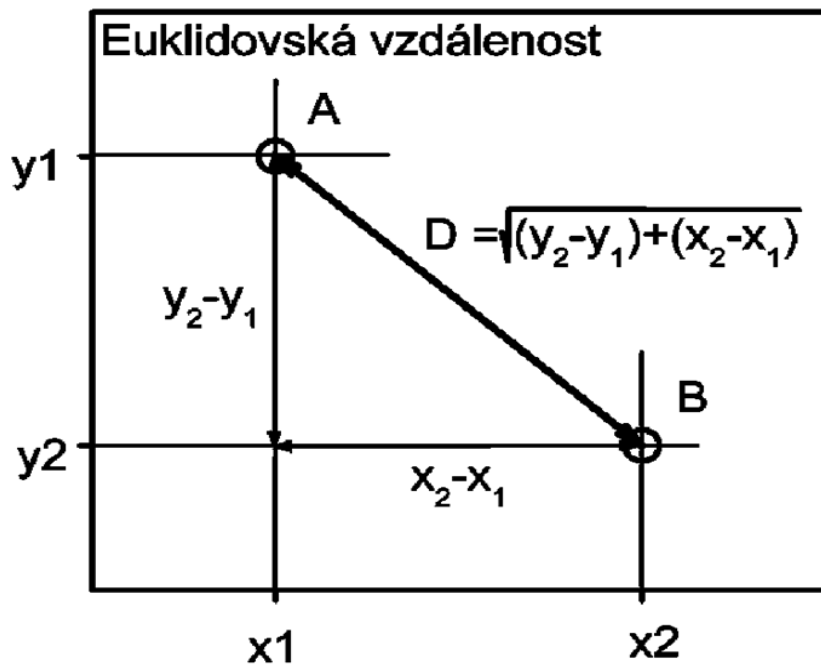
- 1) Eliminuje vychýlení, kvůli rozdílu v lišících se proměnných (různé stupnice, různé jednotky).
- 2) Proměnné se v jednotné stupnici snadno porovnávají. (Kladné hodnoty jsou nad průměrem a záporné hodnoty jsou pod průměrem).
- 3) Změnou stupnice nedojde k rozdílu mezi hodnotami.

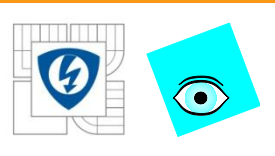


2.3 Míry podobnosti

Podobnost je měřena rozličnými způsoby

Míry vzdálenosti: nejčastěji užívané míry podobnosti. Vzdálenost je reciproká hodnota podobnosti. Čím větší hodnota vzdálenosti, tím menší podobnost.





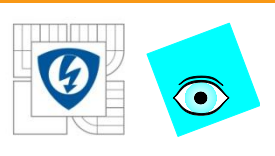
Vzdálenosti

Euklidovská vzdálenost zvaná také **geometrická metrika**

$$d_E(x_k, x_l) = \sqrt{\sum_{j=1}^m (x_{kj} - x_{lj})^2}$$

Manhattanská vzdálenost zvaná také **vzdálenost městských bloků** nebo **Hammingova metrika** je definovaná

$$d_H(x_k, x_l) = \sum_{j=1}^m |x_{kj} - x_{lj}|$$

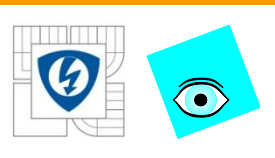


Zobecněná minkovského metrika

Zobecněná Minkovského metrika

$$d_M(x_k, x_l) = \sqrt[z]{\sum_{j=1}^m |x_{kj} - x_{lj}|^z}$$

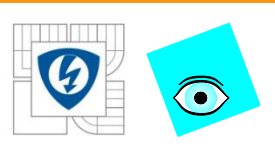
kde pro $z=1$ jde o Hammingovu metriku a pro $z=2$ o Euklidovu. Čím je větší, tím více je zdůrazňován rozdíl mezi vzdálenými objekty.



Tětivová vzdálenost

Tětivová vzdálenost (anglicky chord distance) je definovaná

$$d_{CH}(x_k, x_l) = \sqrt{2 \left[1 - \frac{\sum_{j=1}^m x_{kj} x_{lj}}{\sum_{j=1}^m x_{kj}^2 \sum_{j=1}^m x_{lj}^2} \right]}$$

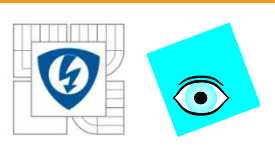


Mahalanobisova metrika

Mahalanobisova metrika pro silně korelované znaky x_k a x_l

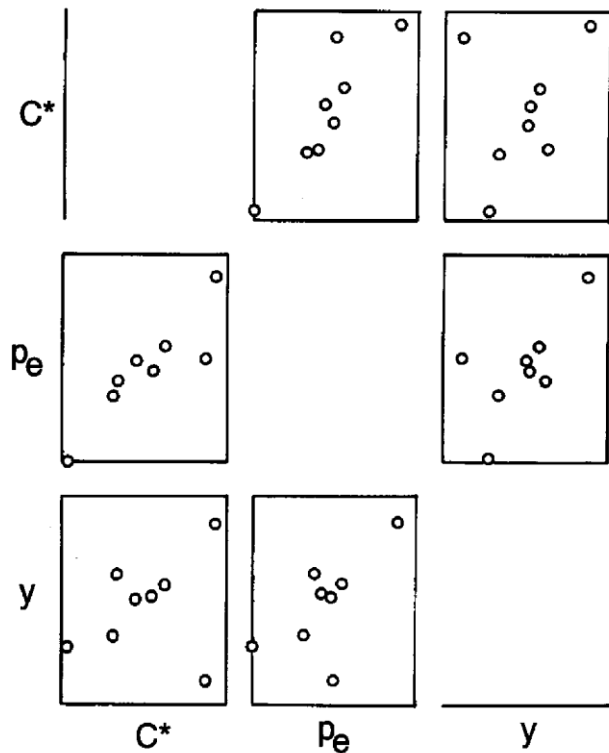
$$d_{Ma}(x_k, x_l) = \sqrt{(x_k - x_l)^T C^{-1} (x_k - x_l)}$$

vyjadřuje vzdálenost bodů v prostoru, jehož osy nemusí být orthogonální.

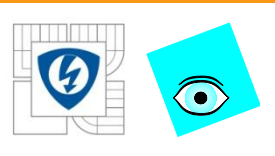


2.4 Korelační míry

Korelační koeficient mezi párem objektů pro několik proměnných.



Vysoká korelace značí vysokou podobnost,
nízká korelace značí nepodobnost.



2.5 Míry asociace

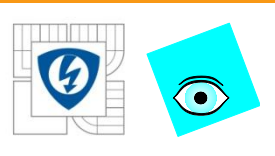
Slouží k porovnání objektů když jejich vlastnosti (znaky) jsou nemetrické (tj. nominální nebo ordinální proměnné).

Asociace mezi dvěma objekty O_i a O_j má možné binární odezvy typu 0-1 v kontingenční tabulce

	Objekt O_i		
Objekt O_j		1	0
	1	a	b
	0	c	d

Všechny možné kombinace počtu znaků pro dva objekty:

- a** značí počet znaků, kde mají oba objekty O_i a O_j hodnotu 1 a jde o tzv. pozitivní shodu.
- b** značí počet znaků, kde mají oba objekty O_i hodnotu 1 a O_j hodnotu 0.
- c** značí počet znaků, kde mají oba objekty O_i hodnotu 0 a O_j hodnotu 1.
- d** značí počet znaků, kde mají oba objekty O_i a O_j hodnotu 0 a jde o tzv. negativní shodu.



Míra asociace

Míra asociace vyjadřují relativní podíly počtu znaků s ohledem na to, zda má smysl uvažovat negativní shodu nebo zda má nulová hodnota znaku u porovnávaných objektů stejnou příčinu.

Sokalův-Michenerův koeficient asociace (čili koeficient jednoduché shody)

$$S_{SM} = \frac{a+d}{a+b+c+d}$$

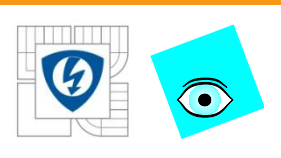
Russelův-Raovův koeficient asociace $S_{RR} = \frac{d}{a+b+c+d}$

Hammanův koeficient asociace $S_H = \frac{a+d-b-c}{a+b+c+d}$

Korelační koeficient $r_{BB} = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$

Rogersův a Tanimotův koeficient asociace $S_{RT} = \frac{a+d}{a+2b+2c+d}$

Sorensenův koeficient asociace $S_S = \frac{2a}{2a+b+c}$



Postup analýzy shluků

Poskytuje empirické a objektivní metody ke klasifikaci objektů

1. krok: Cíle analýzy shluků

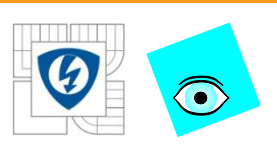
2. krok: Formulace úlohy analýzy shluků

3. krok: Předpoklady analýzy shluků

4. krok: Výstavba dendrogramu shluků

5. krok: Interpretace shluků

6. krok: Validace a profilování shluků



3. krok: Předpoklady analýzy shluků

Analýza shluků není charakteru statistického testování.

Objektivní kvantifikace strukturních vlastností souboru objektů.

Nemá požadavky normality, linearity, homoskedasticity.

Existují pouze dva kritické předpoklady:

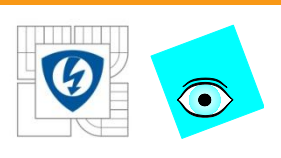
Reprezentativnost vzorku

Výběr objektů a odvozené shluky představují strukturu populace. Zvolený výběr dat musí být opravdovým představitelem populace. Odlehlé objekty zdůrazní divergentní shluky, které zanesou vychýlení do odhadu struktury objektů. Výběr musí být dostatečně reprezentativní a výsledky zobecnitelné na celou populaci.

Vliv multikolinearity

Multikolineární proměnné jsou implicitně váženy intenzivněji. Vyšetřit proměnné na přítomnost multikolinearity:

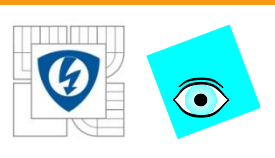
- (1) Je třeba zredukovat počet proměnných
- (2) Použít Mahalanobisovu vzdálenost.



Postup analýzy shluků

Poskytuje empirické a objektivní metody ke klasifikaci objektů

1. krok: Cíle analýzy shluků
2. krok: Formulace úlohy analýzy shluků
3. krok: Předpoklady analýzy shluků
- 4. krok: Výstavba dendrogramu shluků**
- 5. krok: Interpretace shluků**
- 6. krok: Validace a profilování shluků**



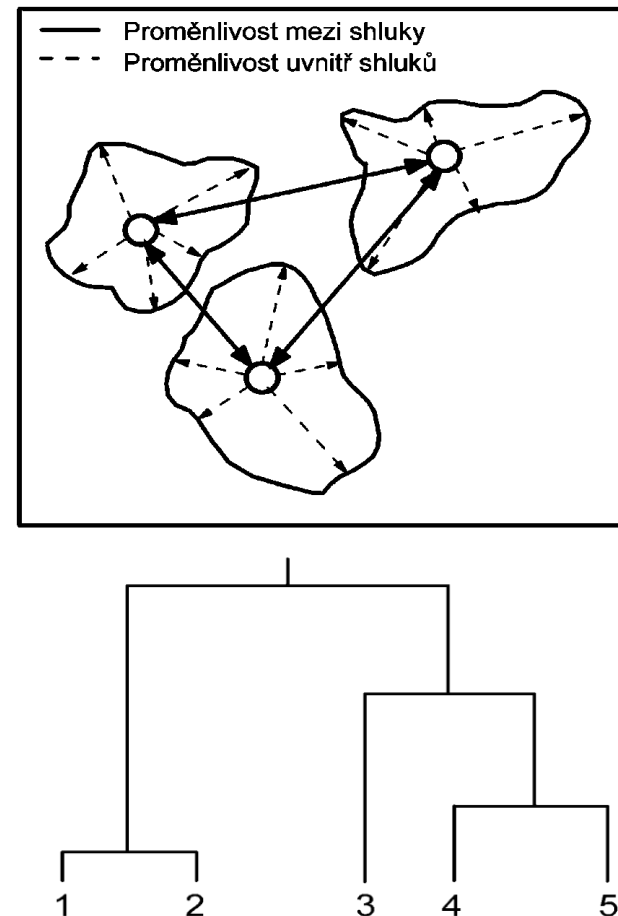
4. krok: Výstavba dendrogramu shluků

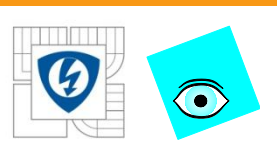
Vedle algoritmu je třeba vybrat i vhodný postup.

Rozlišovací kritérium: maximalizace rozdílů mezi shluky, Proměnlivost mezi shluky vůči proměnlivosti uvnitř shluků.

Test: poměr rozptylu mezi shluky vůči průměru rozptylu uvnitř shluků

Algoritmy se dělí: hierarchické a nehierarchické.





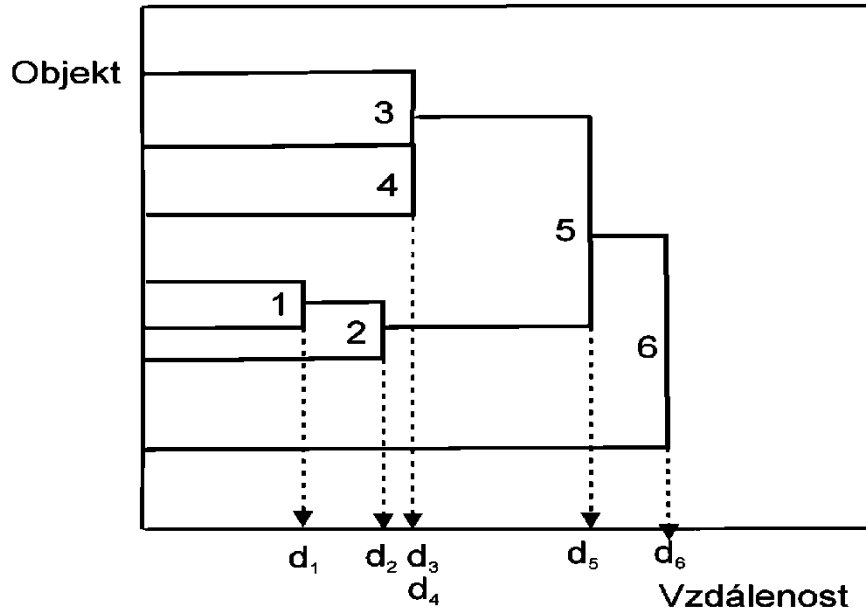
Hierarchické shlukování

konstrukce stromovité struktury, dendrogramu

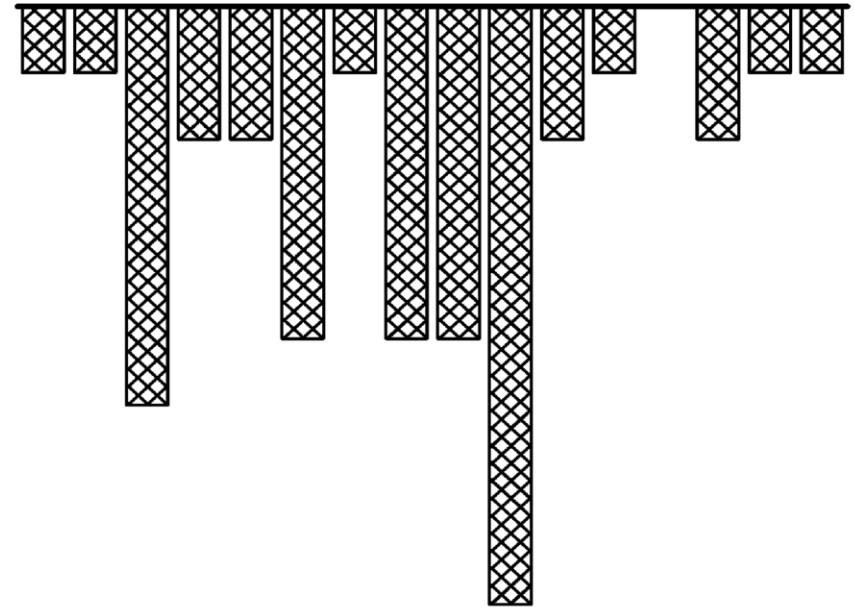
Způsoby hierarchického shlukování: aglomerační a divizní,

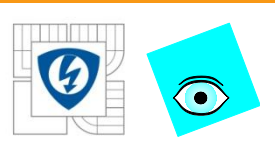
Aglomerační způsob: nejprve se spojí dva nejbližší objekty v jediný shluk, pak se připojí třetí objekt k prvním dvěma objektům a vznikne společný shluk. Tak se seskupí všechny objekty do jednoho velikého shluku.

1) **růstový strom (dendrogram),**
diagram,



2) **vertikální krápníkový**





Aglomerační způsoby (algoritmy) výstavby dendrogramu shluků:

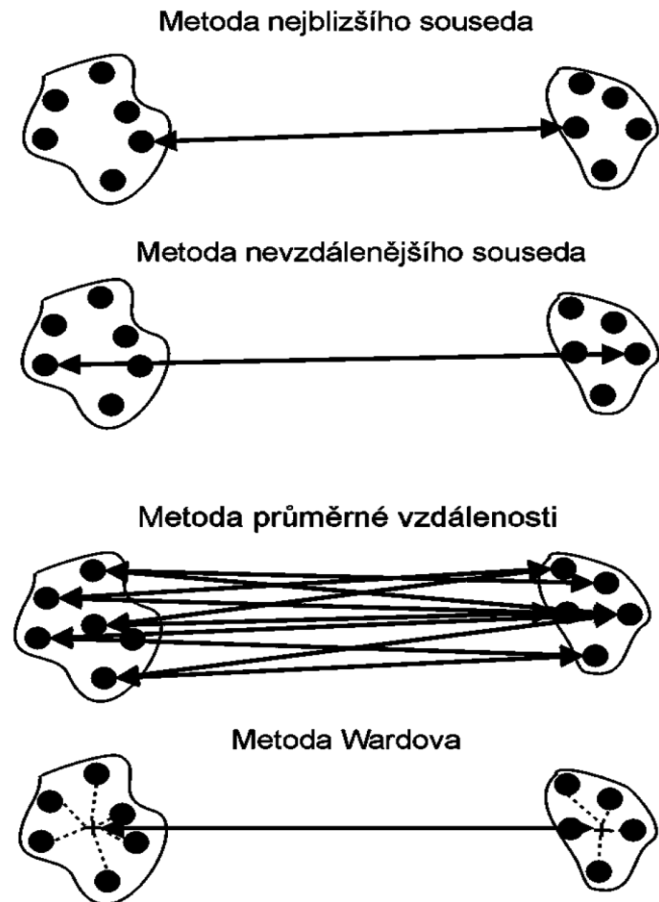
Metoda nejbližšího souseda: je postavena na minimální vzdálenosti objektů.

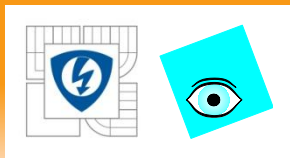
Metoda nejvzdálenějšího souseda: je postavena nikoliv na minimální ale na maximální vzdálenosti.

Metoda průměrového linkování: kritériem je průměrná vzdálenost všech objektů v jednom shluku ke všem objektům ve druhém shluku.

Wardova metoda: vzdálenost mezi dvěma shluky je tvořena na základě sumy čtverců přes všechny proměnné mezi dvěma shluky.

Metoda těžiště: vzdálenost těžišť shluků spojených Euklidovskou vzdáleností nebo čtvercem Euklidovské vzdálenosti. Těžiště shluku je průměrná hodnota objektů v proměnných, vyjádřená ve shlukových proměnných.





Míra věrohodnosti „nejlepšího dendrogramu“

1. kritérium těsnosti proložení:

kofenetický korelační koeficient CC

- nejlépe odpovídá struktuře objektů a znaků mezi objekty,
- je to Pearsonův korelační koeficient mezi skutečnou a predikovanou vzdáleností, založenou na dendrogramu.

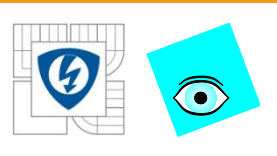
2. kritérium těsnosti proložení

kritérium delta Δ

- měří stupeň přetvoření struktury dat,
- je žádoucí, aby hodnoty delta byly blízké nule,
- je definováno

$$\Delta_A = \left[\frac{\sum_{j < k}^N |d_{jk} - d_{jk}^*|^{\frac{1}{A}}}{\sum_{j < k}^N (d_{jk}^*)^{\frac{1}{A}}} \right]^A$$

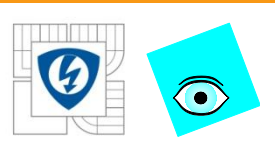
Kde $A=0.5$ nebo 1 , d_{jk} je vzdálenost v původní matici vzdáleností a d_{jk}^* je vzdálenost získaná v dendrogramu.



Úloha 2. Vytvoření dendrogramu objektů neuroleptika (Kompendium B402)

Liší se v účincích: potlačují nervozitu, záchvaty, třes, ospalost, parkinsonismus, vynechávání menstruace, vyrážky, zvýšené slinění. Provedeme klasifikaci neuroleptik do shluků podobných účinků s ohledem na 4 znaky.

Data: Charakter proměnných (převrácená hodnota mediánové účinné dávky 1/ED50 [kg/mg]): **B402x1** značí název neuroleptika, **B402x2** je pro potlačení nervozity, **B402x3** značí potlačení stereotypního chování, **B402x4** je pro potlačení záchvatu a třesu, a **B402x5** znamená dávku smrtícího účinku.



Data

B402x1	B402x2	B402x3	B402x4	B402x5
1 Chlorpromazine	3.846	3.333	1.111	1.923
2 Promazine 0.323	0.213	0.108	1.429	
3 Trifluoperazine	27.027	17.857	0.562	0.14
4 Fluphenazine	17.857	15.385	1.695	1.075
5 Perphenazine	27.027	27.027	1.961	2.083
6 Thioridazine	0.244	0.185	0.093	1.333
7 Pifluthixol	142.857	142.857	20.408	163.934
8 Thiothixene	4.348	4.348	0.047	0.345
9 Chorprothixene	5.882	2.941	4.545	4.167
10 Spiperone 62.5	47.619	11.765	0.847	
11 Haloperidol	52.632	62.5	1.282	0.568
12 Azaperone	2.941	1.282	2.222	3.03
13 Pipamperone	0.327	0.187	1.724	0.397
14 Pimozide 20.408	20.408	0.107	0.025	
15 Metitepine	15.385	10.204	10.204	27.027
16 Clozapine 0.161	0.093	0.327	0.323	
17 Perlazine 0.323	0.323	0.37	0.067	
18 Sulpiride 0.047	0.047	0.003	0.001	
19 Butaclamol	10.204	9.091	1.471	0.025
20 Molindone	7.692	7.692	0.14	0.006

25.2.2010

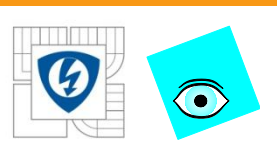
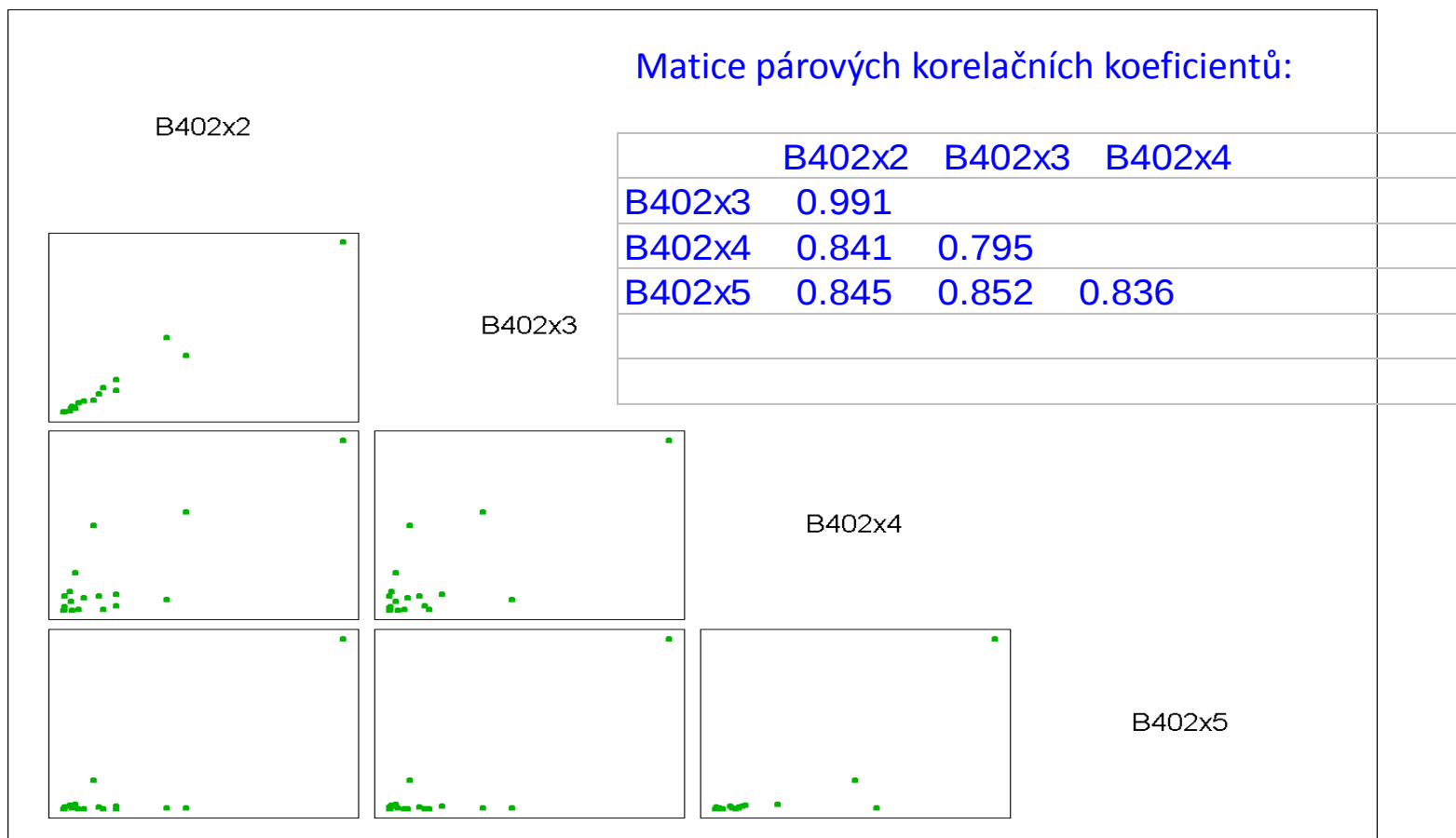


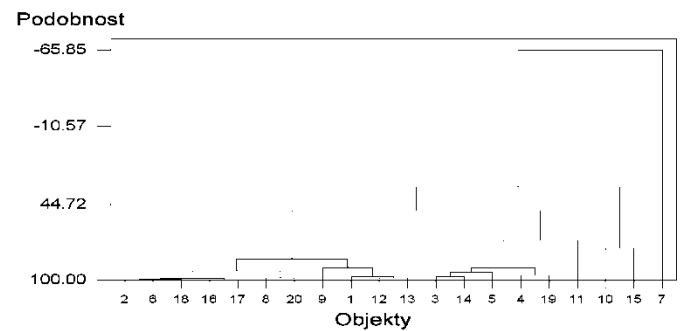
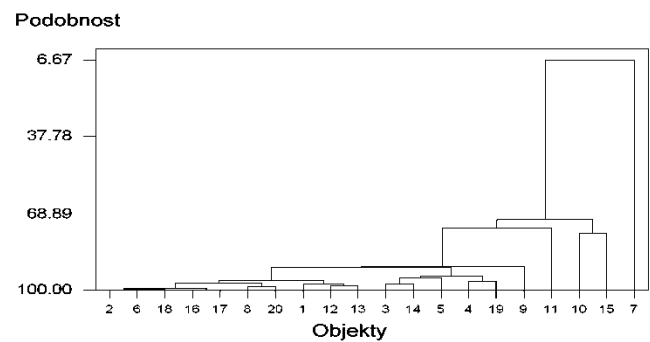
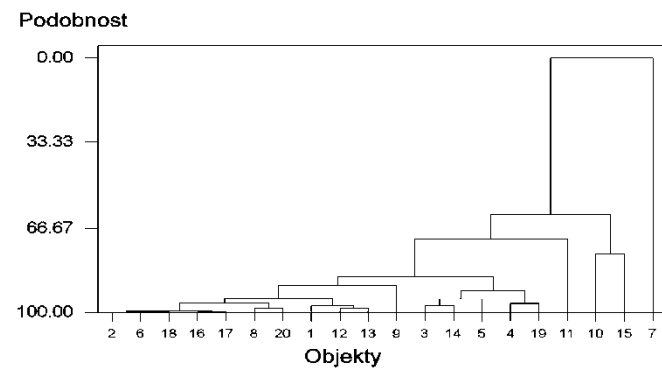
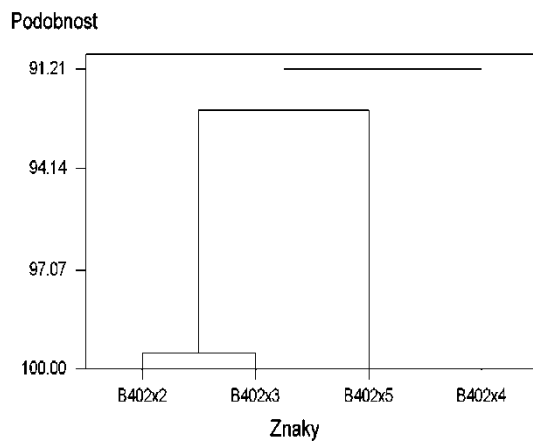
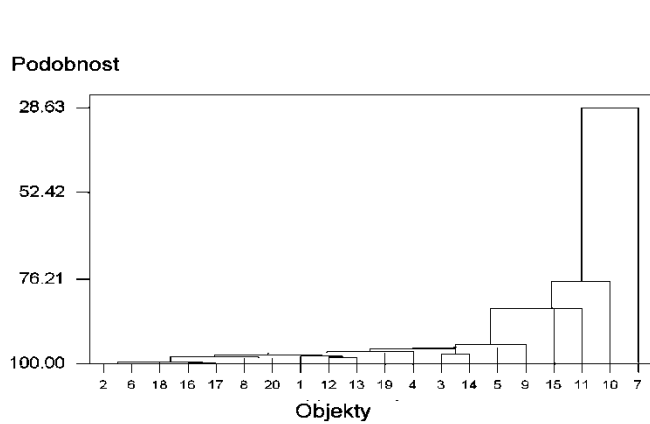
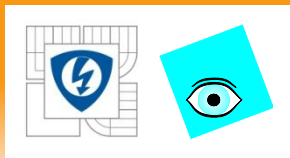
Diagram korelační matice

a statistická významnost korelace pomocí Pearsonových párových korelačních koeficientů





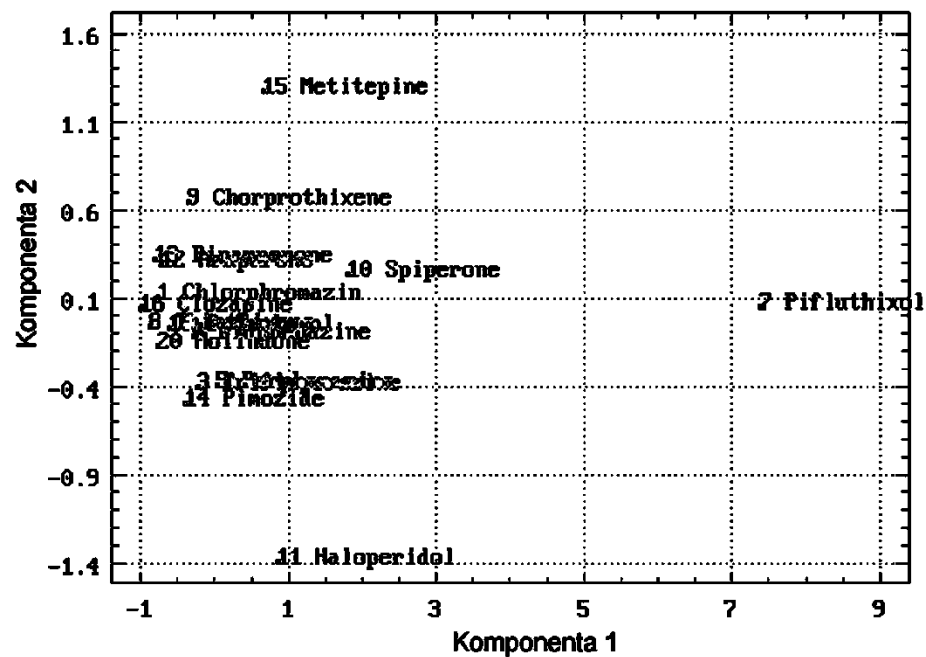
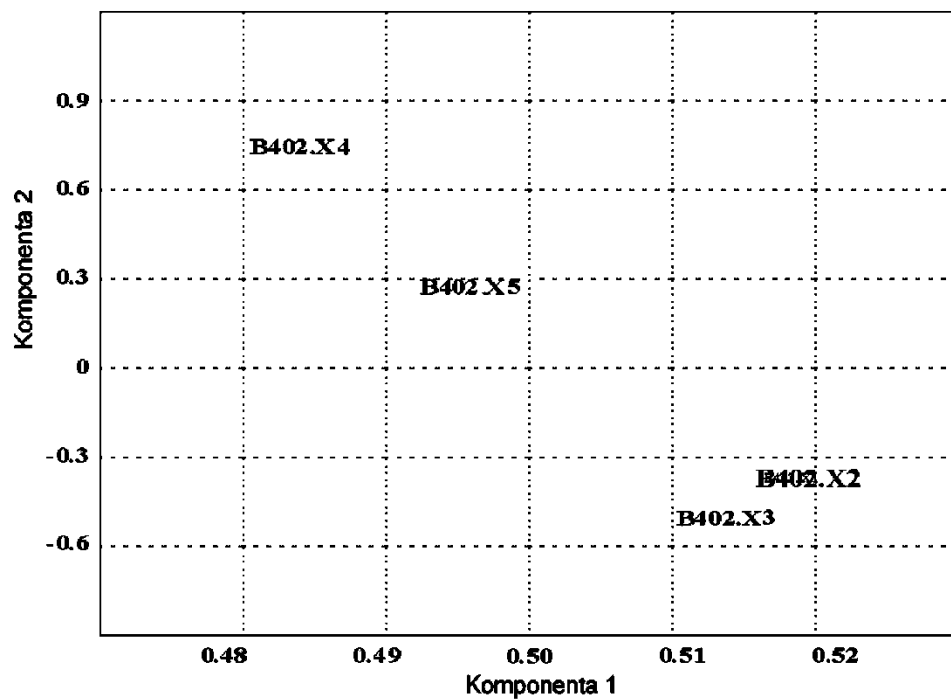
1. Metoda shlukování: **Skupinový průměr**, Typ vzdálenosti: Eucleid., směrodatná odchylka, Kofenetická korelace: **0.987356**, Delta(0.5): **0.137455**, Delta(1.0): **0.125290**;
2. Metoda shlukování: **Jednoduchý průměr**, Typ vzdálenosti: Eucleid., směrodatná odchylka, Kofenetická korelace: **0.988876**, Delta(0.5): **0.177810**, Delta(1.0): **0.188781**;
3. Metoda shlukování: **Těžiště**, Typ vzdálenosti: Eucleid., směrodatná odchylka, Kofenetická korelace: **0.984750**, Delta(0.5): **0.175238**, Delta(1.0): **0.166599**;
4. Metoda shlukování: **Nejbližšího souseda**, Typ vzdálenosti: Eucleid., směrodatná odchylka, Kofenetická korelace: **0.988598**, Delta(0.5): **0.474238**, Delta(1.0): **0.391993**;
5. Metoda shlukování: **Median**, Typ vzdálenosti: Eucleid., směrodatná odchylka, Kofenetická korelace: **0.984215**, Delta(0.5): **0.452308**, Delta(1.0): **0.428346**;
6. Metoda shlukování: **Wardova metoda**, Typ vzdálenosti: Eucleid., směrodatná odchylka, Kofenetická korelace: **0.979285**, Delta(0.5): **0.549394**, Delta(1.0): **0.492716**.

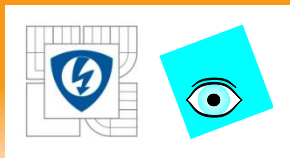


25.2.2010

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ



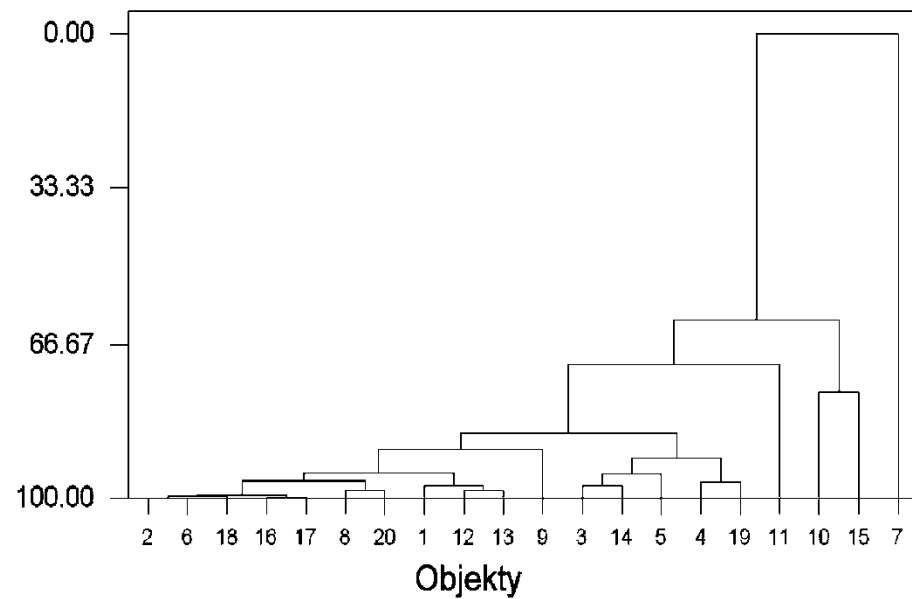




Podobnost



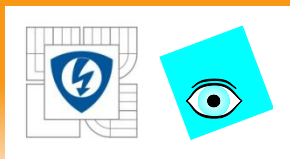
Podobnost



25.2.2010

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ



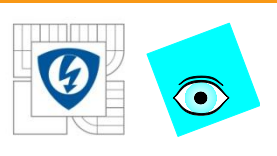


Nehierarchické shlukování

netýká se výstavby stromu, objekty se přidělují do shluků, když je počet shluků předem zadán.

Postup:

1. krok: **Zadání zárodku** shluku (= počátečního středu shluku).
2. krok: Objekty **uvnitř zadané vzdálenosti** budou do shluku zařazeny.
3. krok: Zvolen **zárodek jiného shluku** a zařazování pokračuje.
4. krok: Existuje několik postupů **K-means shlukování** (nejbližších středů, těžišť):
 - (a) **Sekvenční práh**: začíná volbou jednoho zárodku a zahrnuje všechny objekty uvnitř předspecifikované vzdálenosti. Když jsou všechny zahrnuty, je vybrán zárodek druhého shluku, atd.
 - (b) **Paralelní práh**: vybírá několik zárodků současně (paralelně) a zařazuje objekty uvnitř prahové vzdálenosti do nejbližšího zárodku.
 - (c) **Optimalizace**: dovoluje znovuzařazení objektů. Když se objekt octne blíže jinému shluku, než se právě nachází, optimalizační postup ho přeřadí do jiného, bližšího shluku.

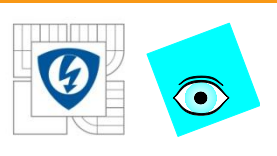


Volba zárodků shluků v nehierarchickém shlukování

Sekvenční prahový postup je ukázkou pro velké datové soubory.

Postup: zadá se počet shluků a začnou se vybírat zárodky shluků dle kroků:

1. Prvním zárodkem je první úplný objekt zdrojové matice dat,
2. Druhým zárodkem je další úplný objekt, který je oddělen od prvního zadanou minimální vzdáleností.
3. Po zadání všech zárodků začne zařazování objektů.
4. Klíčovým problémem zůstává volba shlukových zárodků.



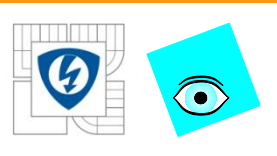
Hierarchické nebo nehierarchické metody?

(1) Užívání hierarchických metod:

- a) Hierarchické metody mají výhodu, že jsou rychlé.
- b) Hierarchické metody mohou být klamné, pro nežádoucí předešlé shluky setrvávající v průběhu analýzy.
- c) Působení odlehlých objektů, proto vypouštět jen velmi opatrně.
- d) Hierarchické metody nejsou stavěné na analýzu velmi velikých výběrů.

(2) Užívání nehierarchických metod:

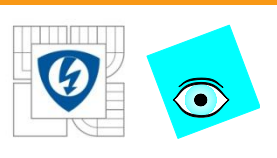
- a) V poslední době se nehierarchické metody využívají stále více.
- b) Použití nehierarchické metody závisí na schopnosti uživatele, jeho praktických zkušenostech a objektivní teorii jak vybrat zárodkové body.
- c) Výsledky nehierarchické metody jsou méně ovlivněny odlehlými body.
- d) U nehierarchické metody se užívají vzdálenostní míry a lze užít i nepodstatné proměnné.
- e) U nehierarchické metody lze pouze za použití nenáhodných zárodkových bodů.



Hierarchické nebo nehierarchické metody?

(3) Kombinace obou metod, hierarchických a nehierarchických:

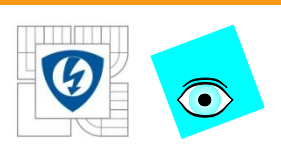
- a) Nejprve hierarchickou metodou určíme: počet shluků, profily shlukovaných center a zřetelné odlehlé body.
- b) Po odstranění odlehlých bodů: zbývající objekty shlukovány nehierarchicky se zárodky z výsledků hierarchické metody.



Počet vytvářených shluků

Terminační kritérium provádí vyšetření podobnosti mezi shluky po každém kroku, a to **když míra podobnosti překročí předdefinovanou velikost nebo když následné hodnoty se skokově změní.**

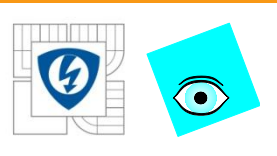
Doporučení: určí se rozličný počet shluků např. 2, 3 a 4 a na základě praktického úsudku se pak rozhodne.



Postup analýzy shluků

Poskytuje empirické a objektivní metody ke klasifikaci objektů

1. krok: Cíle analýzy shluků
2. krok: Formulace úlohy analýzy shluků
3. krok: Předpoklady analýzy shluků
4. krok: Výstavba dendrogramu shluků
5. krok: Interpretace shluků
6. krok: Validace a profilování shluků

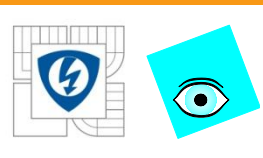


5. krok: Interpretace shluků

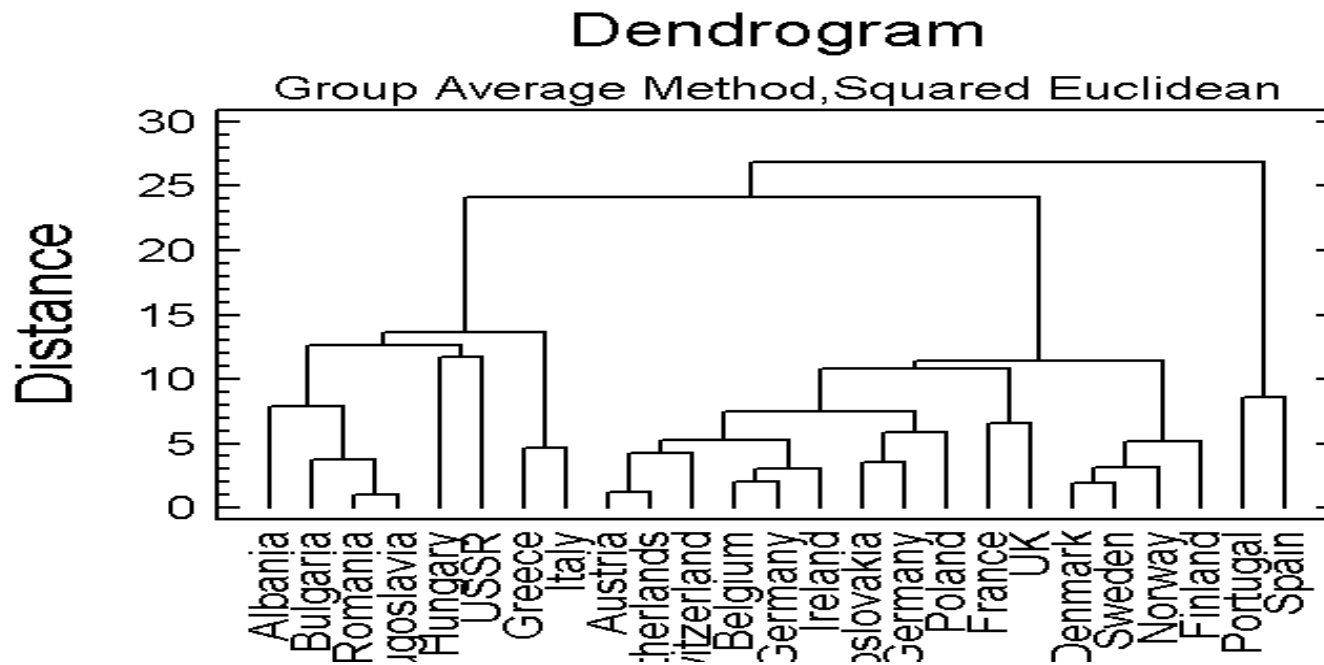
- (a) Vyšetření každého shluku v pojmech shlukových proměnných.
- (b) Pojmenování shluků nebo jeho označení, které vystihuje jeho podstatu a povahu.

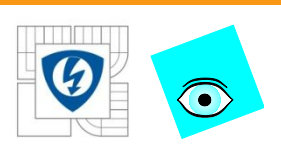
Profilování a interpretace shluků:

- (a) Prokazuje popis.
- (b) Přidělení korespondence ke shlukům předvídaným z teorie.
- (c) V konfirmatorním modu profily přidělují shlukům korespondenci.
- (d) Při hledání korespondence nebo praktické významnosti by se měly porovnávat odvozené shluky s předem vytvořenou typologií.



Dendrogram

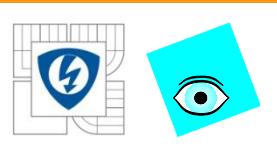




Postup analýzy shluků

Poskytuje empirické a objektivní metody ke klasifikaci objektů

1. krok: Cíle analýzy shluků
2. krok: Formulace úlohy analýzy shluků
3. krok: Předpoklady analýzy shluků
4. krok: Výstavba dendrogramu shluků
5. krok: Interpretace shluků
- 6. krok: Validace a profilování shluků**



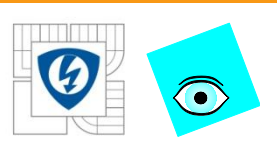
6. krok: Validace a profilování shluků

Existuje subjektivní charakter hledání optimálního shlukového řešení.
Neexistuje jednoduchá metoda, která by zajišťovala validitu a praktický význam.

Validování shluků: znamená, že nalezené shlukové řešení

- (a) je reprezentativní,
- (b) je zobecnitelné na ostatní objekty v celém původním souboru,
- (c) je stabilní i v čase.

Postup: - analyzovat oddělené výběry,
- porovnat nalezená shluková řešení a
- odhadnout shodu výsledků.



6. krok: Validace a profilování shluků

Rozdělení výběru dat na dva vzorky: každý vzorek je podroben analýze shluků odděleně a výsledky jsou porovnány:

- (1) Modifikovanou formu rozdělení výběru, kdy v prvním vzorku získáme středy shluků a využijeme je k definování shluků ve druhém vzorku objektů a výsledky porovnáme,
- (2) Přímá forma vzájemného porovnání (cross-validation).

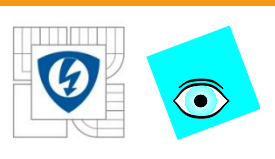
Způsob vytyčení kritéria: Užijeme takové proměnné, které sice nejsou užity k vytvoření shluků, ale mění se dostatečně od shluku ke shluku.



Úloha 3. Sledování spotřeby proteinů v Evropě (Kompendium B418)

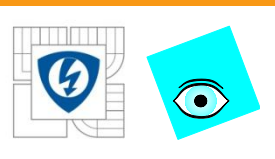
Sledovaná spotřeba proteinů v 25 zemích formou spotřeby 9 druhů potravin je předmětem vyšetření.

Data: *i* značí index, **Cervene** udává červené maso, **Bile** maso, **Vejce**, **Mleko**, **Ryby**, **Obilniny**, **Skrob**, **Orechy**, **Ovoce** a zelenina



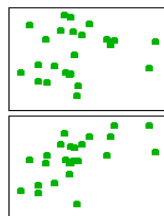
Úloha 3. Sledování spotřeby proteinů v Evropě (Kompendium B418)

	Objekty	Proměnné								
i	Stát	Cervene	Bile	Vejce	Mleko	Ryby	Obilniny	Skrob	Orechy	Ovoce
1	Albania	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
2	Austria	8.9	14	4.3	19.9	2.1	28	3.6	1.3	4.3
3	Belgium	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4
4	Bulgaria	7.8	6	1.6	8.3	1.2	56.7	1.1	3.7	4.2
5	Czechoslov.	9.7	11.4	2.8	12.5	2	34.3	5	1.1	4
6	Denmark	10.6	10.8	3.7	25	9.9	21.9	4.8	0.7	2.4
7	E Germany	8.4	11.6	3.7	11.1	5.4	24.6	6.5	0.8	3.6
8	Finland	9.5	4.9	2.7	33.7	5.8	26.3	5.1	1	1.4
9	France	18	9.9	3.3	19.5	5.7	28.1	4.8	2.4	6.5
10	Greece	10.2	3	2.8	17.6	5.9	41.7	2.2	7.8	6.5
11	Hungary	5.3	12.4	2.9	9.7	0.3	40.1	4	5.4	4.2
12	Ireland	13.9	10	4.7	25.8	2.2	24	6.2	1.6	2.9
13	Italy	9	5.1	2.9	13.7	3.4	36.8	2.1	4.3	6.7
14	Netherlands	9.5	13.6	3.6	23.4	2.5	22.4	4.2	1.8	3.7
15	Norway	9.4	4.7	2.7	23.3	9.7	23	4.6	1.6	2.7
16	Poland	6.9	10.2	2.7	19.3	3	36.1	5.9	2	6.6
17	Portugal	6.2	3.7	1.1	4.9	14.2	27	5.9	4.7	7.9
18	Romania	6.2	6.3	1.5	11.1	1	49.6	3.1	5.3	2.8
19	Spain	7.1	3.4	3.1	8.6	7	29.2	5.7	5.9	7.2
20	Sweden	9.9	7.8	3.5	24.7	7.5	19.5	3.7	1.4	2
21	Switzerland	13.1	10.1	3.1	23.8	2.3	25.6	2.8	2.4	4.9
22	UK	17.4	5.7	4.7	20.6	4.3	24.3	4.7	3.4	3.3
23	USSR	9.3	4.6	2.1	16.6	3	43.6	6.4	3.4	2.9
24	W Germany	11.4	12.5	4.1	18.8	3.4	18.6	5.2	1.5	3.8
25	Yugoslavia	4.4	5	1.2	9.5	0.6	55.9	3	5.7	3.2

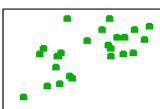


Test významnosti korelace v korelační matici

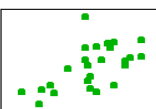
Cervene



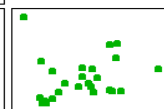
Bile



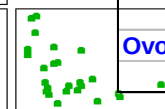
Vejce



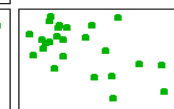
Mleko



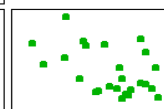
Ryby



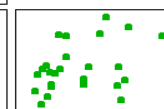
Obilniny



Skrob



Orechy

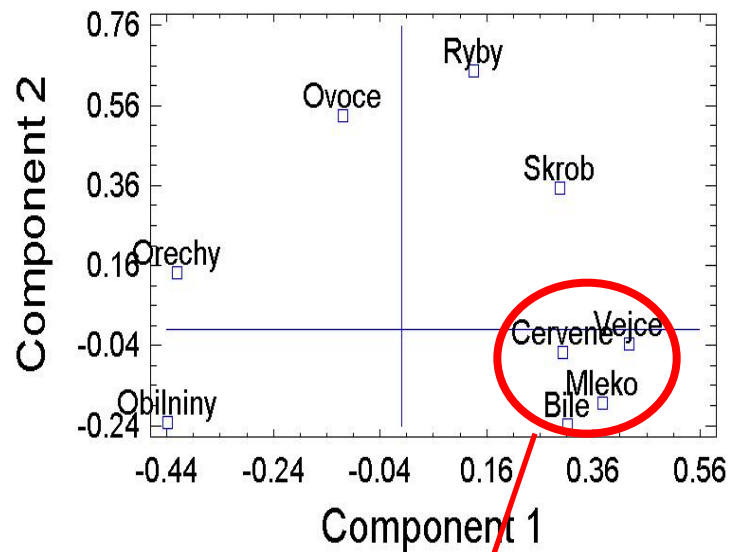


Ovoce

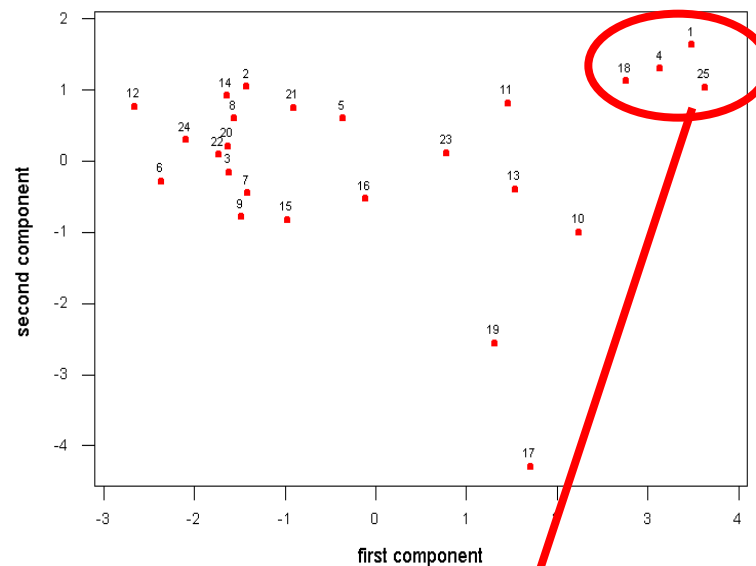
	Cervene	Bile	Vejce	Mleko	Ryby	Obilniny	Skrob	Orechy	Ovoce
Cervene									
Bile	0.153								
Vejce	0.5856	0.6204							
Mleko	0.0021	0.0009							
Ryby	0.5029	0.2815	0.5755						
Obilniny	0.0104	0.1728	0.0026						
Skrob	0.061	-0.234	0.0656	0.1379					
Orechy	0.7722	0.2602	0.7555	0.511					
Ovoce	-0.4999	-0.4138	-0.7124	-0.5927	-0.5242				
	0.0109	0.0398	0.0001	0.0018	0.0071				
	0.1354	0.3138	0.4522	0.2224	0.4039	-0.5333			
	0.5186	0.1267	0.0232	0.2853	0.0453	0.0061			
	-0.3494	-0.635	-0.5598	-0.6211	-0.1472	0.651	-0.4743		
	0.0869	0.0007	0.0036	0.0009	0.4827	0.0004	0.0166		
	-0.0742	-0.0613	-0.0455	-0.4084	0.2661	0.0465	0.0844	0.375	
	0.7244	0.7709	0.829	0.0427	0.1985	0.8251	0.6883	0.0648	

Pod hodnotami korelačních koeficientů jsou vypočtené hladiny významnosti P. Je-li $P < 0.05$, je korelace významná.

Plot of Component Weights

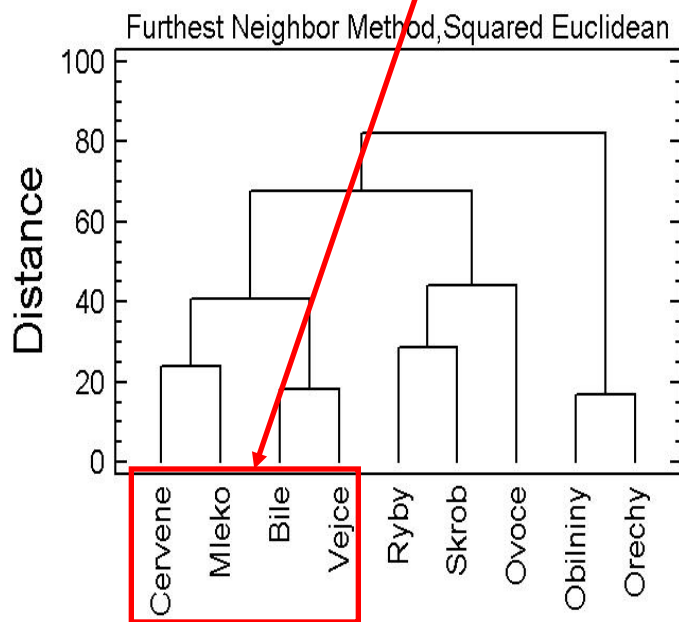


Principal Components Score Plot

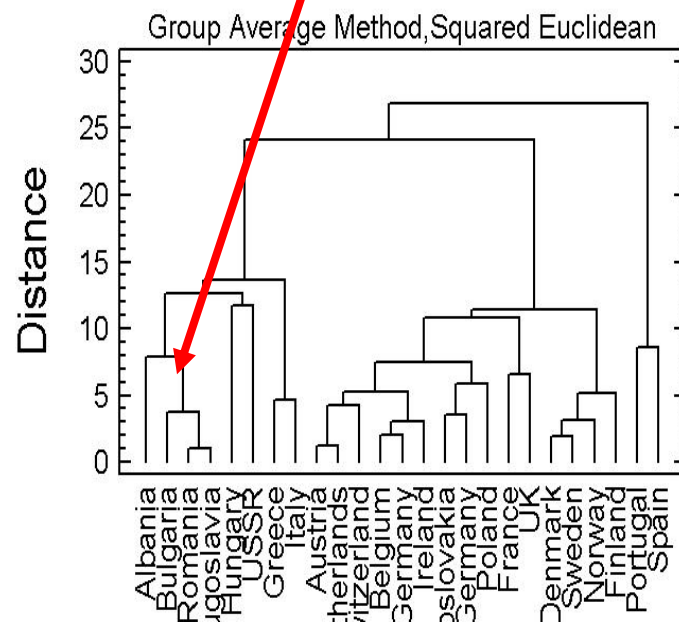


- 1 Albania
- 2 Austria
- 3 Belgium
- 4 Bulgaria
- 5 Czechoslovakia
- 6 Denmark
- 7 E Germany
- 8 Finland
- 9 France
- 10 Greece
- 11 Hungary
- 12 Ireland
- 13 Italy
- 14 Netherlands
- 15 Norway
- 16 Poland
- 17 Portugal
- 18 Romania
- 19 Spain
- 20 Sweden
- 21 Switzerland
- 22 UK
- 23 USSR
- 24 W Germany
- 25 Yugoslavia

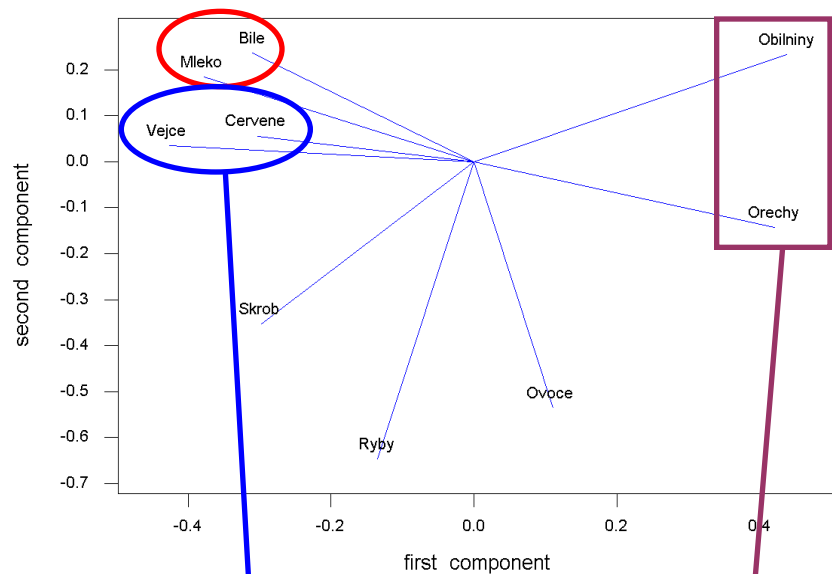
Dendrogram



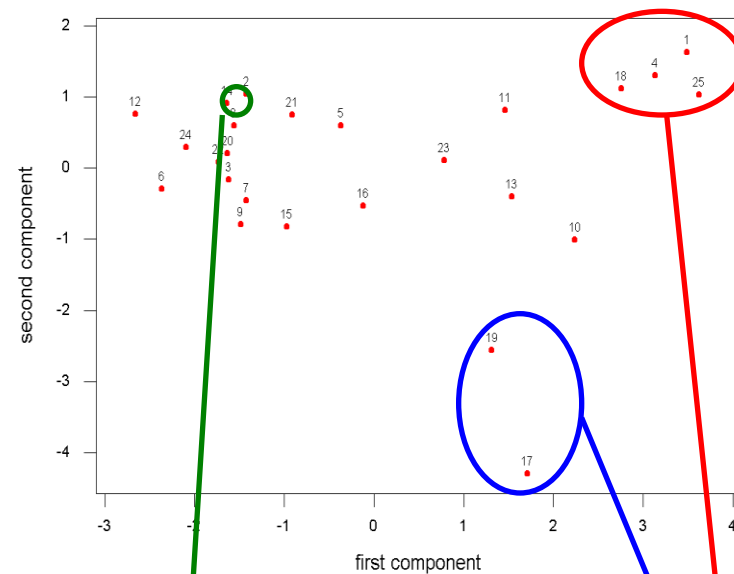
Dendrogram



Principal Components Loading Plot



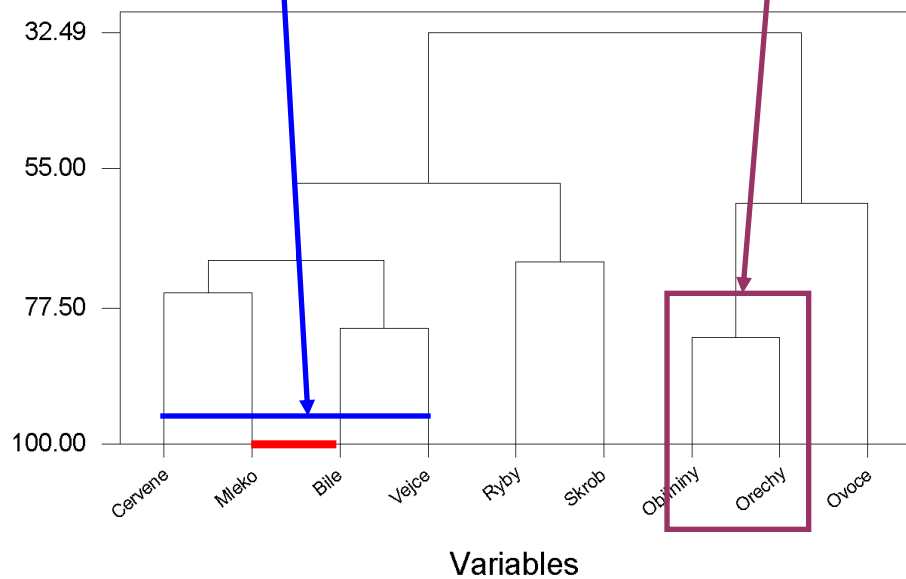
Principal Components Score Plot



- 1 Albania
- 2 Austria
- 3 Belgium
- 4 Bulgaria
- 5 Czechoslovakia
- 6 Denmark
- 7 E Germany
- 8 Finland
- 9 France
- 10 Greece
- 11 Hungary
- 12 Ireland
- 13 Italy
- 14 Netherlands
- 15 Norway
- 16 Poland
- 17 Portugal
- 18 Romania
- 19 Spain
- 20 Sweden
- 21 Switzerland
- 22 UK
- 23 USSR
- 24 W Germany
- 25 Yugoslavia

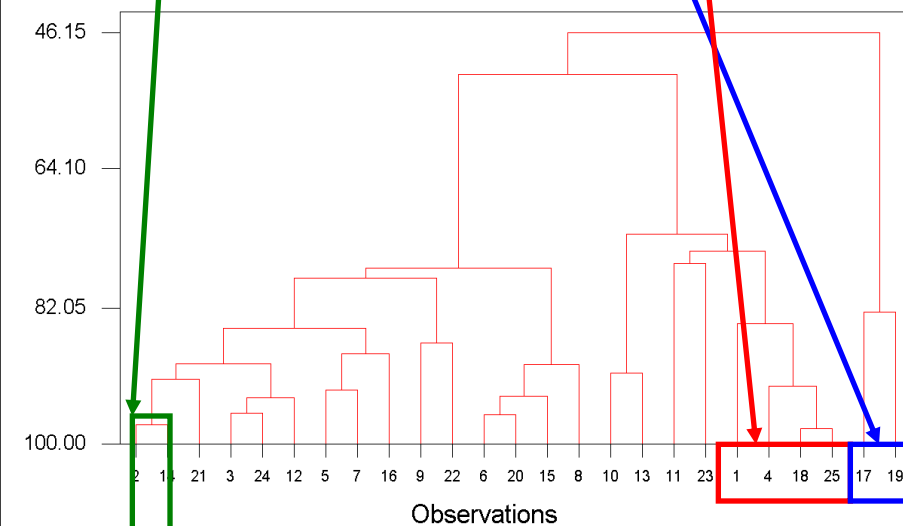
Average, Euclid., Standard.,

Similarity



Dendrogram

Similarity



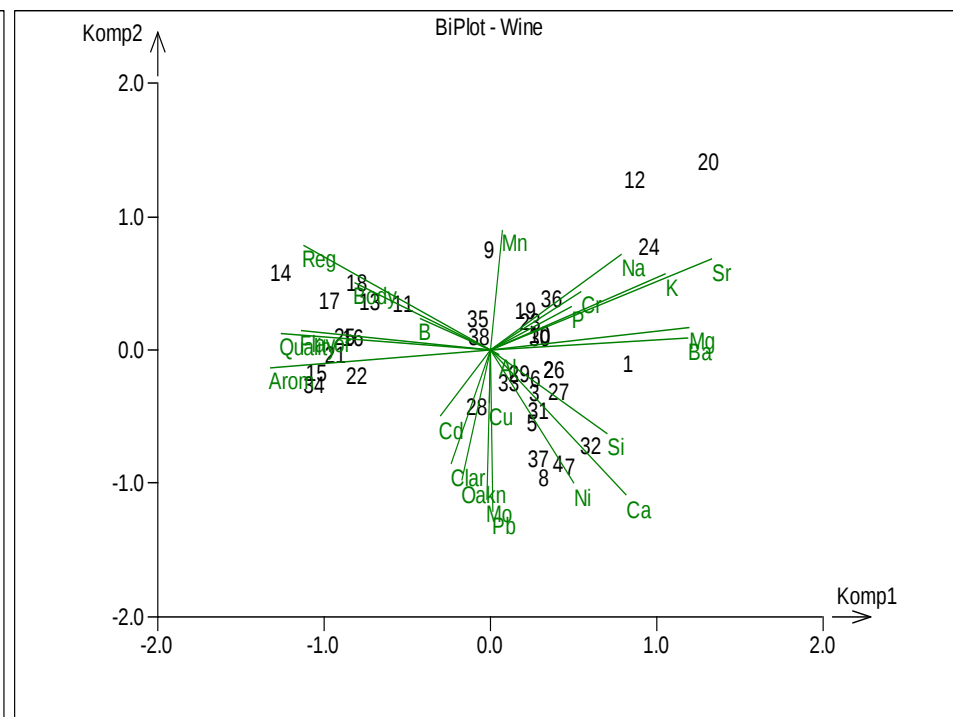
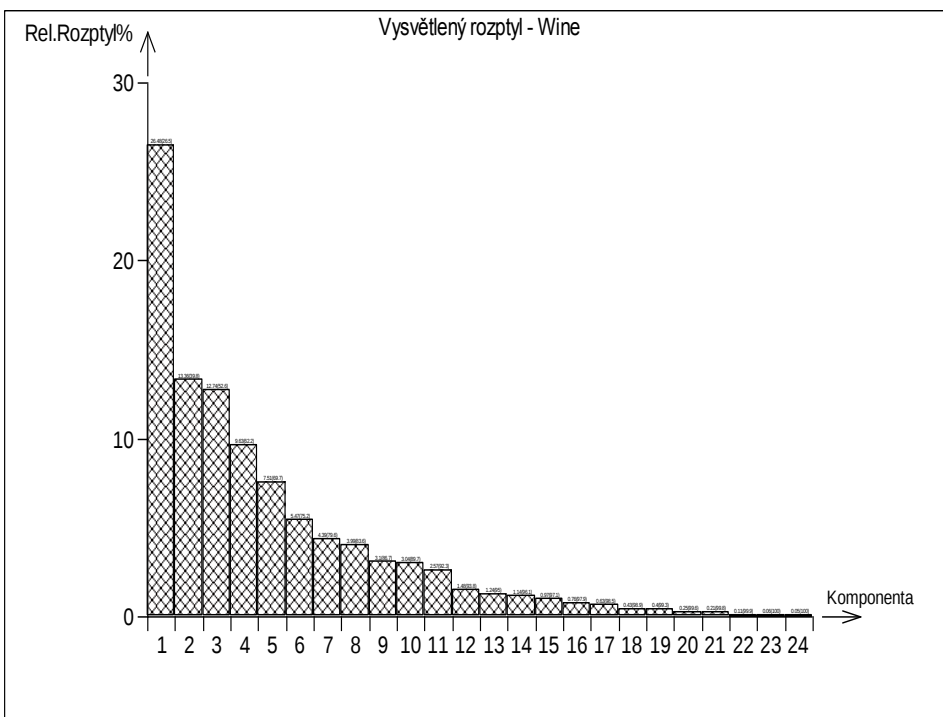
Úloha 4. Faktorová analýza při klasifikaci vzorků vín (Kompendium E408)

Pro 38 vzorků vín bylo nalezeno 24 analytických obsahů stopových prvků a charakteristických fyzikálně-chemických vlastností. Utvořte shluky podobných vlastností a dále shluky podobných vín.

Index	Cd	Mo	Mn	Ni	Cu	Al	Ba	Cr	Sr	Pb	B	Mg	Si	Na	Ca	P	K	Arom	Clar	Body	Flavor	Oakn	Quality	Reg
1	0.005	0.044	1.51	0.122	0.83	0.982	0.387	0.029	1.23	0.561	2.63	128	17.3	66.8	80.5	150	1130	3.3	1	2.8	3.1	4.1	9.8	1
2	0.055	0.16	1.16	0.149	0.066	1.02	0.312	0.038	0.975	0.697	6.21	193	19.7	53.3	75	118	1010	4.4	1	4.9	3.5	3.9	12.6	1
3	0.056	0.146	1.1	0.088	0.643	1.29	0.308	0.035	1.14	0.73	3.05	127	15.8	35.4	91	161	1160	3.9	1	5.3	4.8	4.7	11.9	1
4	0.063	0.191	0.959	0.38	0.133	1.05	0.165	0.036	0.927	0.796	2.57	112	13.4	27.5	93.6	120	924	3.9	1	2.6	3.1	3.6	11.1	1
5	0.011	0.363	1.38	0.16	0.051	1.32	0.38	0.059	1.13	1.73	3.07	138	16.7	76.6	84.6	164	1090	5.6	1	5.1	5.5	5.1	13.3	1
6	0.05	0.106	1.25	0.114	0.055	1.27	0.275	0.019	1.05	0.491	6.56	172	18.7	15.7	112	137	1290	4.6	1	4.7	5	4.1	12.8	1
7	0.025	0.479	1.07	0.168	0.753	0.715	0.164	0.062	0.823	2.06	4.57	179	17.8	98.5	122	184	1170	4.8	1	4.8	4.8	3.3	12.8	1
8	0.024	0.234	0.906	0.466	0.102	0.811	0.271	0.044	0.963	1.09	3.18	145	14.3	10.5	91.9	187	1020	5.3	1	4.5	4.3	5.2	12	1
9	0.009	0.058	1.84	0.042	0.17	1.8	0.225	0.022	1.13	0.048	6.13	113	13	54.4	70.2	158	1240	4.3	1	4.3	3.9	2.9	13.6	3
10	0.033	0.074	1.28	0.098	0.053	1.35	0.329	0.03	1.07	0.552	3.3	140	16.3	70.5	74.7	159	1100	4.3	1	3.9	4.7	3.9	13.9	1
11	0.039	0.071	1.19	0.043	0.163	0.971	0.105	0.028	0.491	0.31	6.56	103	9.5	45.3	67.9	133	1090	5.1	1	4.3	4.5	3.6	14.4	3
12	0.045	0.147	2.76	0.071	0.074	0.483	0.301	0.087	2.14	0.546	3.5	199	9.2	80.4	66.3	212	1470	3.3	0.5	5.4	4.3	3.6	12.3	2
13	0.06	0.116	1.15	0.055	0.18	0.912	0.166	0.041	0.578	0.518	6.43	111	11.1	59.7	83.8	139	1120	5.9	0.8	5.7	7	4.1	16.1	3
14	0.067	0.166	1.53	0.041	0.043	0.512	0.132	0.026	0.229	0.699	7.27	107	6	55.2	44.9	148	854	7.7	0.7	6.6	6.7	3.7	16.1	3
15	0.077	0.261	1.65	0.073	0.285	0.596	0.078	0.063	0.156	1.02	5.04	94.6	6.3	10.4	54.9	132	899	7.1	1	4.4	5.8	4.1	15.5	3
16	0.064	0.191	1.78	0.067	0.552	0.633	0.085	0.063	0.192	0.777	5.56	110	7	13.6	64.1	167	976	5.5	0.9	5.6	5.6	4.4	15.5	3
17	0.025	0.009	1.57	0.041	0.081	0.655	0.072	0.021	0.172	0.232	3.79	75.9	6.4	11.6	48.1	132	995	6.3	1	5.4	4.8	4.6	13.8	3
18	0.02	0.027	1.74	0.046	0.153	1.15	0.094	0.021	0.358	0.025	4.24	80.9	7.9	38.9	57.6	136	876	5	1	5.5	5.5	4.1	13.8	3
19	0.034	0.05	1.15	0.058	0.058	1.35	0.294	0.006	1.12	0.206	2.71	120	14.7	68.1	64.8	133	1050	4.6	1	4.1	4.3	3.1	11.3	1
20	0.013	0.03	2.82	0.058	0.05	0.623	0.349	0.082	2.91	0.171	3.54	208	9.3	79.2	66.4	266	1430	3.4	0.9	5	3.4	3.4	7.9	2
21	0.043	0.268	2.32	0.066	0.314	0.627	0.099	0.045	0.36	1.28	5.68	98.4	9.1	19.5	64.3	176	945	6.4	0.9	5.4	6.6	4.8	15.1	3
22	0.061	0.245	1.61	0.07	0.172	2.07	0.071	0.053	0.186	1.19	4.42	87.6	7.6	11.6	70.6	156	820	5.5	1	5.3	5.3	3.8	13.5	3
23	0.047	0.161	1.47	0.154	0.082	0.546	0.181	0.06	0.898	0.747	8.11	160	19.3	12.5	82.1	218	1220	4.7	0.7	4.1	5	3.7	10.8	2
24	0.048	0.146	1.85	0.092	0.09	0.889	0.328	0.1	1.32	0.604	6.42	134	19.3	125	83.2	173	1810	4.1	0.7	4	4.1	4	9.5	2
25	0.049	0.155	1.73	0.051	0.158	0.653	0.081	0.037	0.164	0.767	4.91	86.5	6.5	11.5	53.9	172	1020	6	1	5.4	5.7	4.7	12.7	3
26	0.042	0.126	1.7	0.112	0.21	0.508	0.299	0.054	0.995	0.686	6.94	129	43.6	45	85.9	165	1330	4.3	1	4.6	4.7	4.9	11.6	2
27	0.058	0.184	1.28	0.095	0.058	1.3	0.346	0.037	1.17	1.28	3.29	145	16.7	65.8	72.8	175	1140	3.9	1	4	5.1	5.1	11.7	1
28	0.065	0.211	1.65	0.102	0.055	0.308	0.206	0.028	0.72	1.02	6.12	99.3	27.1	20.5	95.2	194	1260	5.1	1	4.9	5	5.1	11.9	2
29	0.065	0.129	1.56	0.166	0.151	0.373	0.281	0.034	0.889	0.638	7.28	139	22.2	13.3	84.2	164	1200	3.9	1	4.4	5	4.4	10.8	2
30	0.068	0.166	3.14	0.104	0.053	0.368	0.292	0.039	1.11	0.831	4.71	125	17.6	13.9	59.5	141	1030	4.5	1	3.7	2.9	3.9	8.5	2
31	0.067	0.199	1.65	0.119	0.163	0.447	0.292	0.058	0.927	1.02	6.97	131	38.3	42.9	85.9	164	1390	5.2	1	4.3	5	6	10.7	2
32	0.084	0.266	1.28	0.087	0.071	1.14	0.158	0.049	0.794	1.3	3.77	143	19.7	39.1	128	146	1230	4.2	0.8	3.8	3	4.7	9.1	1
33	0.069	0.183	1.94	0.07	0.095	0.465	0.225	0.037	1.19	0.915	2	123	4.6	7.5	69.4	123	943	3.3	1	3.5	4.3	4.5	12.1	1
34	0.087	0.208	1.76	0.061	0.099	0.683	0.087	0.042	0.168	1.33	5.04	92.9	7	12	56.3	157	949	6.8	1	5	6	5.2	14.9	3
35	0.074	0.142	2.44	0.051	0.052	0.737	0.408	0.022	1.16	0.745	3.94	143	6.8	36.8	67.6	82	1170	5	0.8	5.7	5.5	4.8	13.5	1
36	0.084	0.171	1.85	0.088	0.038	1.21	0.263	0.072	1.35	0.899	2.38	130	6.2	101	64.4	99	1070	3.5	0.8	4.7	4.2	3.3	12.2	1
37	0.106	0.307	1.15	0.063	0.051	0.643	0.29	0.031	0.885	1.61	4.4	151	17.4	7.3	103	177	1100	4.3	0.8	5.5	3.5	5.8	10.3	1
38	0.102	0.342	4.08	0.065	0.077	0.752	0.366	0.048	1.08	1.77	3.37	145	5.3	33.1	58.3	117	1010	5.2	0.8	4.8	5.7	3.5	13.2	1



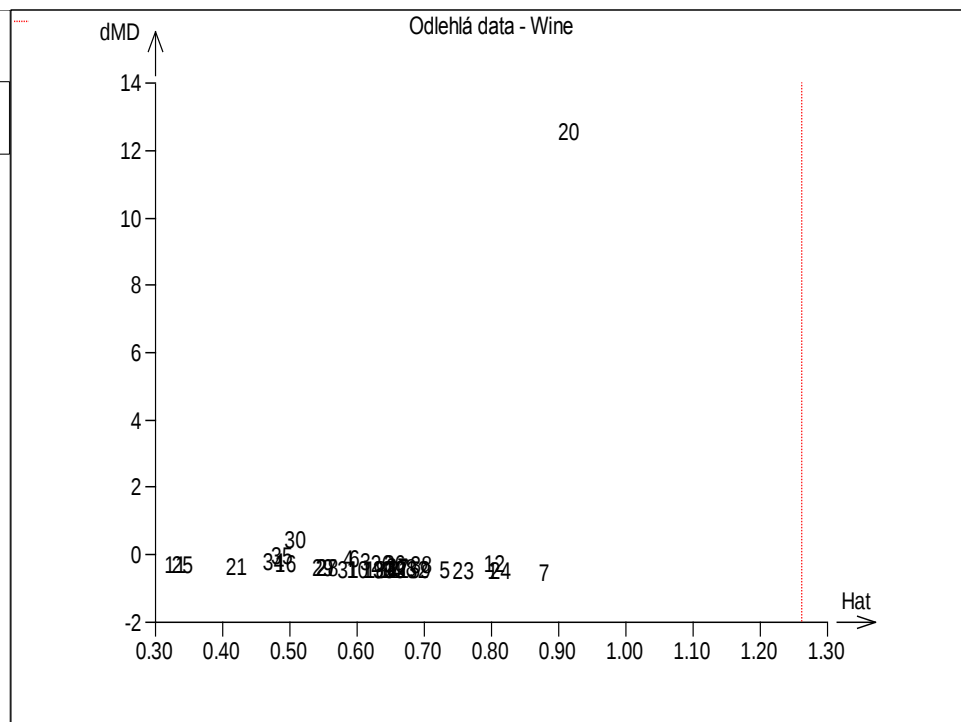
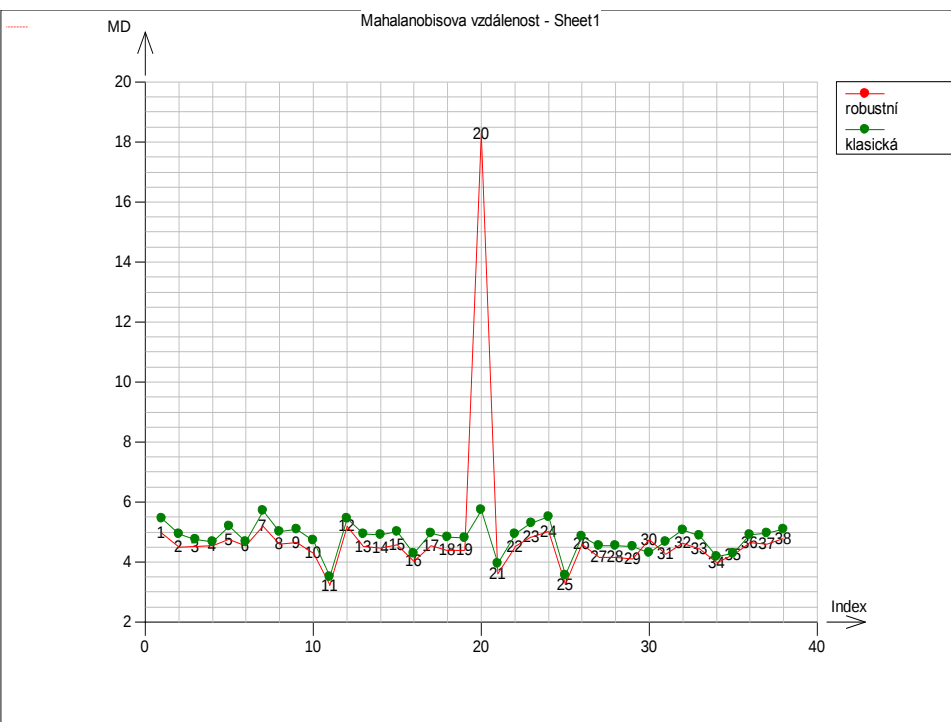
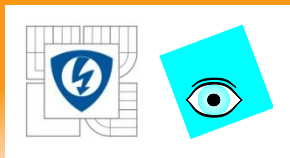
Úloha 4



25.2.2010

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

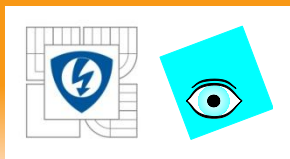




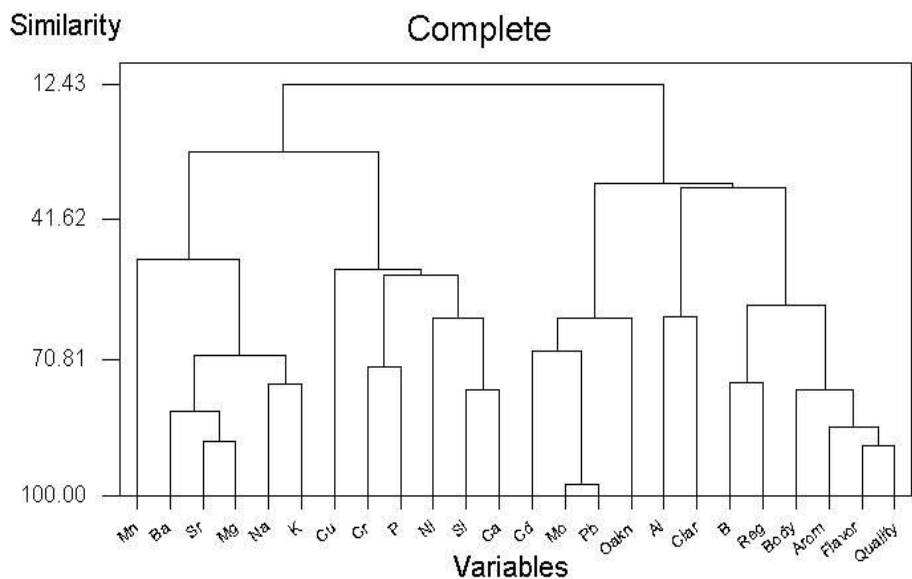
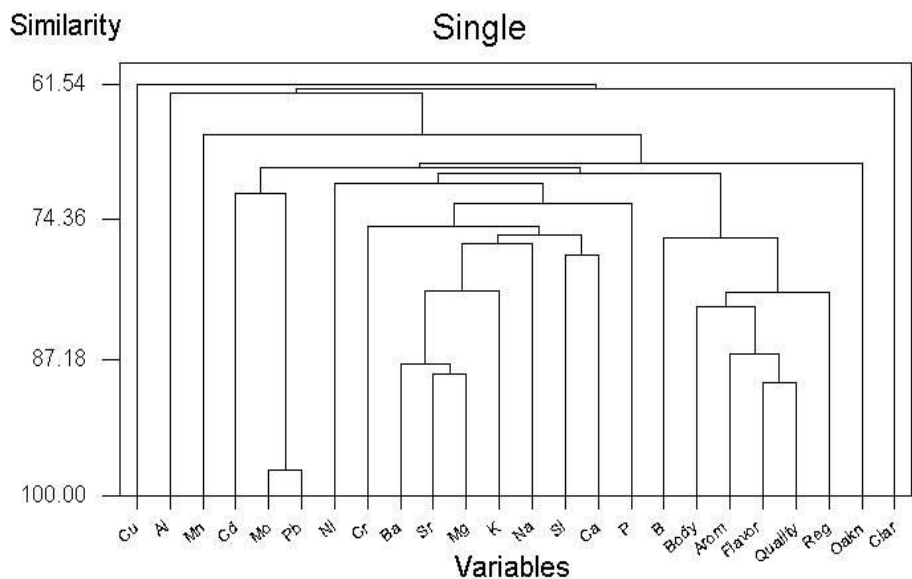
25.2.2010

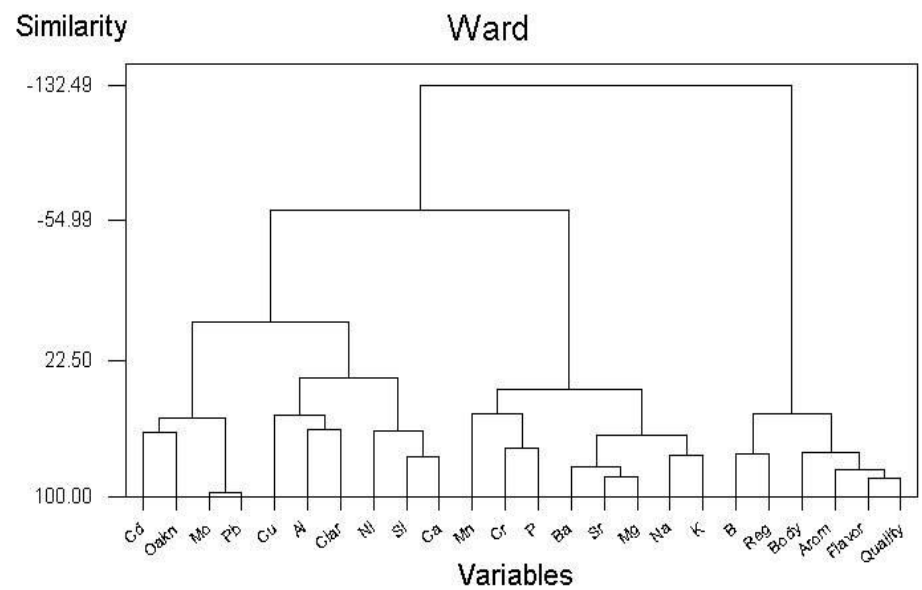
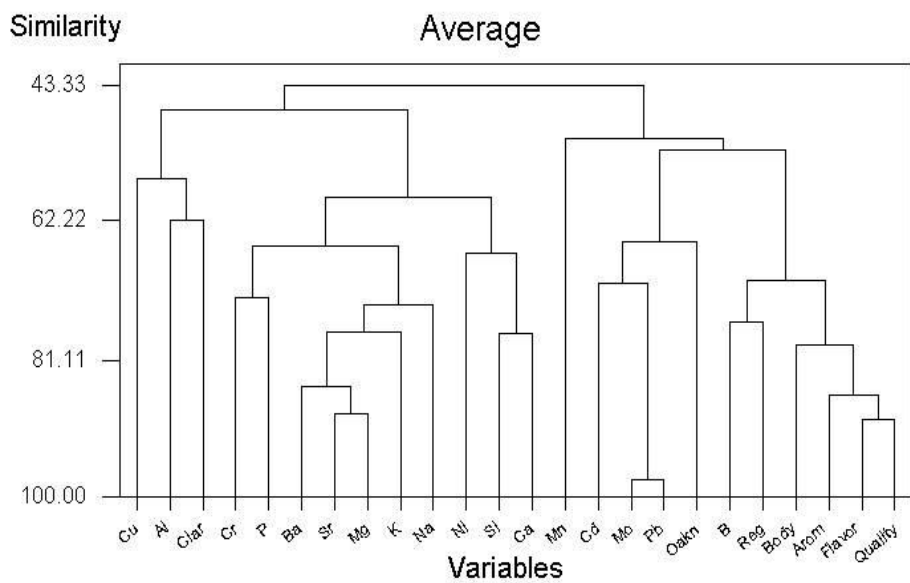
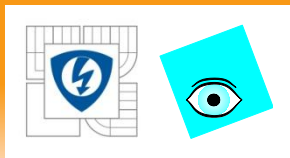
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

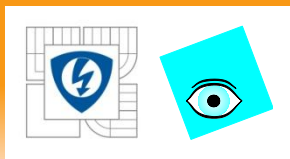




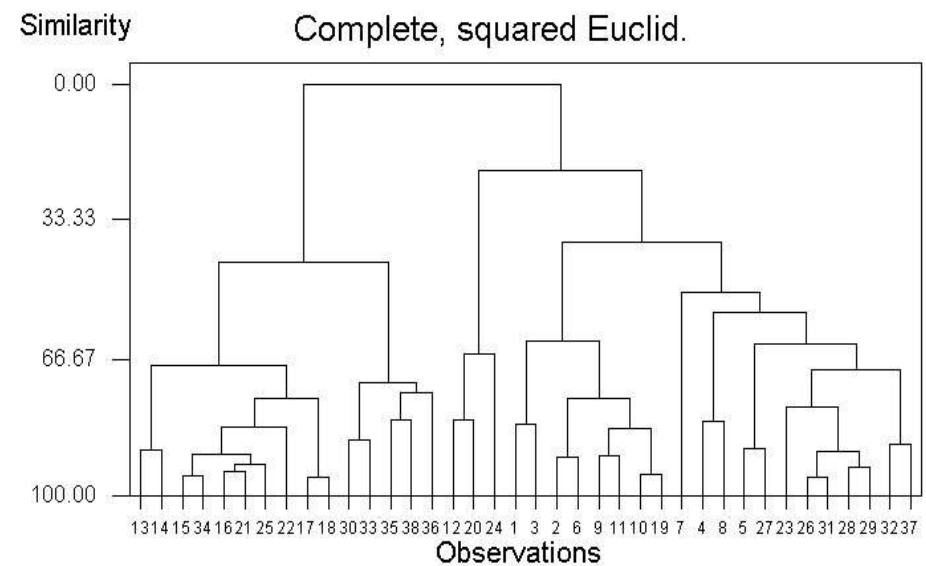
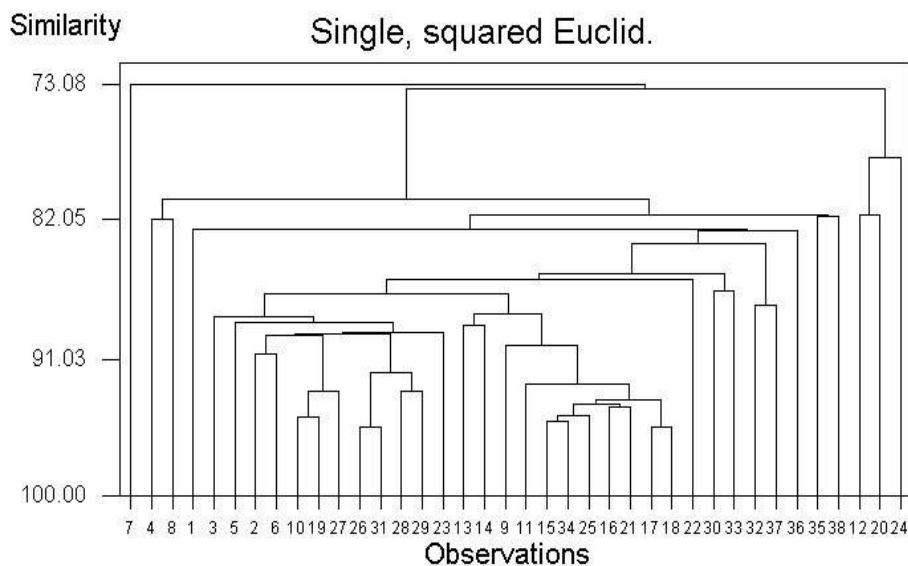
Hledání nejlepší metody dle kofenetického korelačního koeficientu a kritéria delta.







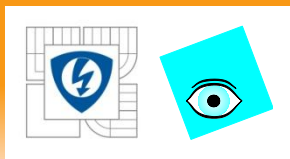
Hledání nejlepší metody dle kofenetického korelačního koeficientu a kritéria delta.



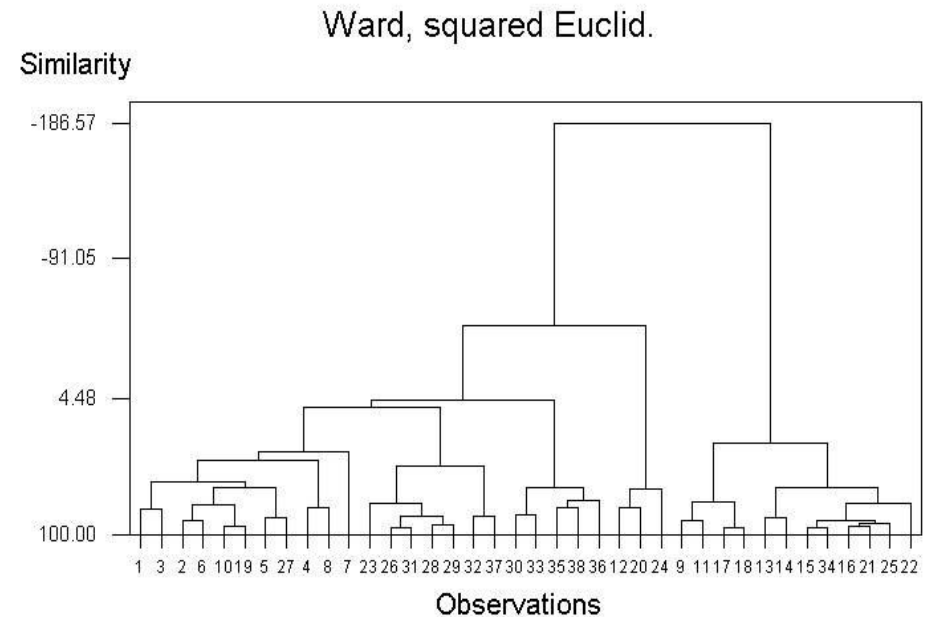
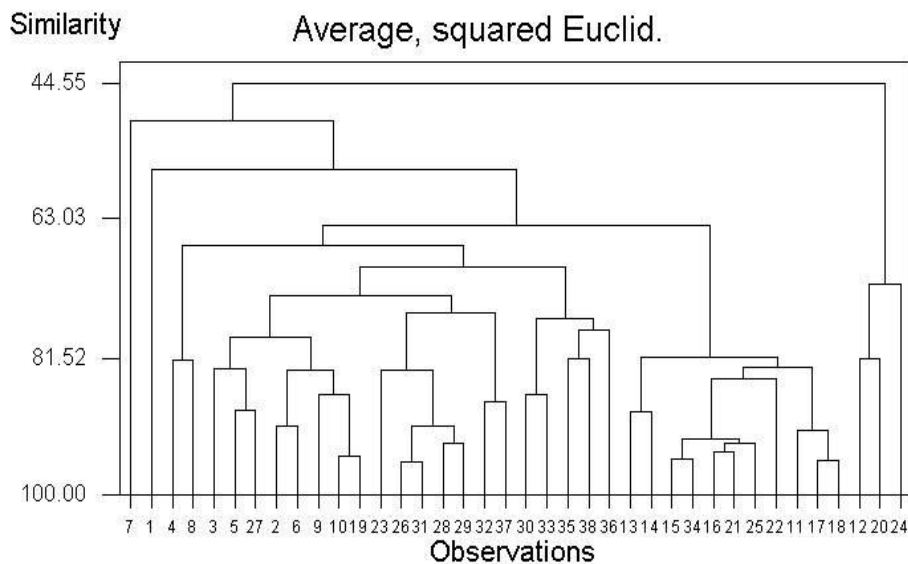
25.2.2010

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ





Hledání nejlepší metody dle kofenetického korelačního koeficientu a kritéria delta.



25.2.2010

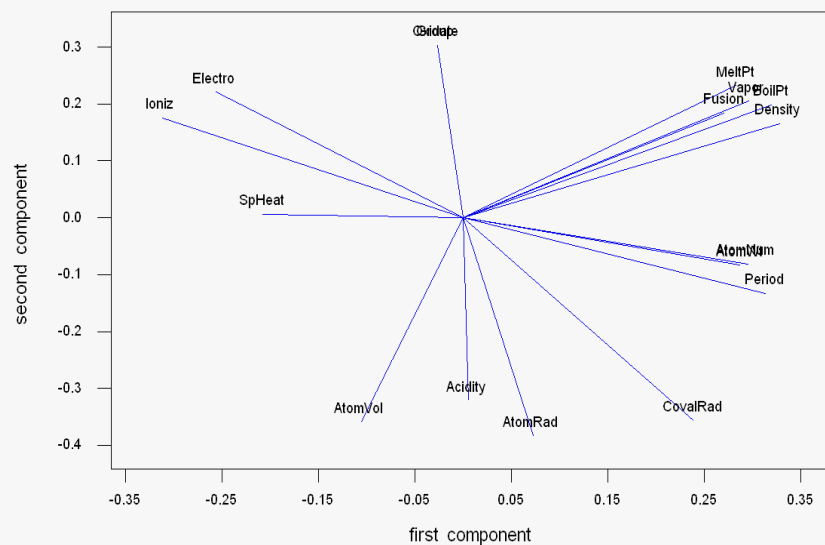
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Úloha 5. Klasifikace prvků periodické tabulky do shluků

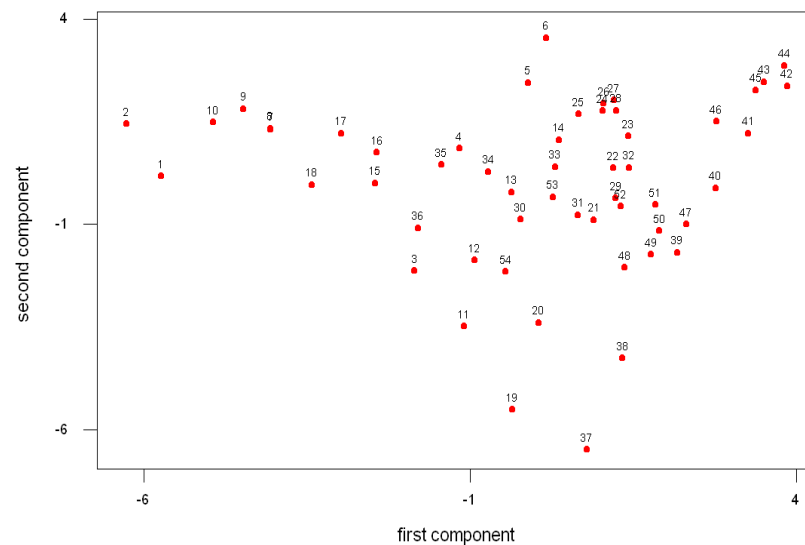
Pro 54 prvků periodické tabulky bylo použito 18 rozličných fyzikálně-chemických vlastností. Naleznete shluky podobných vlastností a shluky podobných prvků.

Row	Elem	At.A.	Per.	A.W.	Ox	Ioniz	Elec.	BoilPt	MeltPt	Vapor	Fusio	SpHeat	AtomVol	CovalRad	AtomRad	Density	Acidity	
1	H	1	1	1	1	313	2.1	20	14	0.11	0.01	3.45	14.1	0.32	0.99	0.07	3	
2	He	2	1	4	8	567	5	4	3	0.02	0.01	1.25	31.8	0.93	1.3	0.13	3	
3	Li	3	2	6.9	1	124	1	1603	454	32.5	0.72	0.79	13.1	1.23	1.55	0.53	5	
4	Be	4	2	9	2	215	1.5	3043	1550	73.9	2.8	0.45	5	0.9	1.12	1.85	3	
5	B	5	2	11	3	191	2	4198	2303	128	5.3	0.309	4.6	0.82	0.98	2.34	2	
6	C	6	2	12	4	260	2.5	5103	4000	172	2.6	0.165	5.3	0.77	0.91	2.26	2	
7	N	7	2	14	5	336	3	77	63	0.67	0.09	0.247	17.3	0.75	0.92	0.81	1	
8	O	8	2	16	6	314	3.5	90	54	0.82	0.05	0.218	14	0.73	0.84	1.14	3	
9	F	9	2	19	7	402	4	85	54	0.76	0.06	0.18	17.1	0.72	0.81	1.5	3	
10	Ne	10	2	20	8	497	5	31	25	0.42	0.08	0.2	16.8	0.71	1.76	1.2	3	
11	Na	11	3	23	1	119	0.9	1165	371	24.1	0.62	0.295	23.7	1.54	1.9	0.97	5	
12	Mg	12	3	24	2	176	1.2	1380	923	32.5	2.14	0.25	14	1.36	1.6	1.74	5	
13	Al	13	3	27	3	138	1.5	2723	933	67.9	2.55	0.215	10	1.18	1.43	2.7	3	
14	Si	14	3	28	4	188	1.8	2953	1683	40.6	11.1	0.162	12.1	1.11	1.32	2.33	3	
15	P	15	3	31	5	254	2.1	553	317	2.97	0.15	0.177	17	1.06	1.28	1.82	2	
16	S	16	3	32	6	239	2.5	718	392	3.01	0.34	0.175	15.5	1.02	1.27	2.07	1	
17	Cl	17	3	36	7	300	3	238	172	2.44	0.77	0.116	18.7	0.99	1.09	1.56	1	
18	Ar	18	3	40	8	363	4	87	84	1.56	0.28	0.125	24.2	0.98	2.11	1.4	3	
19	K	19	4	39	1	100	0.8	1033	337	18.9	0.55	0.177	45.3	2.03	2.35	0.86	5	
20	Ca	20	4	40	2	141	1	1713	1111	36.7	2.1	0.149	29.9	1.74	1.97	1.55	5	
21	Sc	21	4	45	3	151	1.3	3003	1812	81	3.8	0.13	15	1.44	1.62	3	4	
22	Ti	22	4	48	4	158	1.5	3533	1941	107	3.7	0.126	10.6	1.32	1.47	4.51	3	
23	V	23	4	51	5	156	1.6	3723	2173	106	4.2	0.12	8.4	1.22	1.34	6.1	3	
24	Cr	24	4	52	6	156	1.6	2938	2148	73	3.3	0.11	7.2	1.18	1.3	7.19	1	
25	Mn	25	4	55	7	171	1.5	2423	1518	53.7	3.5	0.115	7.4	1.17	1.35	7.43	1	
26	Fe	26	4	56	8	182	1.8	3273	1809	84.6	3.67	0.11	7.1	1.17	1.26	7.86	3	
27	Co	27	4	59	8	181	1.8	3173	1768	93	3.64	0.099	6.7	1.16	1.25	8.9	3	
28	Ni	28	4	59	8	176	1.8	3003	1726	91	4.21	0.105	6.6	1.15	1.24	8.9	4	
29	Cu	29	5	64	1	178	1.9	2868	1356	72.8	3.11	0.092	7.1	1.17	1.28	8.96	4	
30	Zn	30	5	65	2	216	1.6	1179	693	27.4	1.76	0.91	9.2	1.25	1.38	7.14	3	
31	Ga	31	5	70	3	138	1.6	2510	303	70.7	1.34	0.079	11.8	1.26	1.41	5.91	3	
32	Ge	32	5	73	4	187	1.8	3103	1211	68	7.6	0.073	13.6	1.22	1.37	5.32	3	
33	As	33	5	75	5	231	2	886	1090	7.75	6.62	0.083	13.1	1.2	1.39	5.72	2	
34	Se	34	5	79	6	225	2.4	958	490	3.34	1.25	0.084	16.5	1.16	1.4	4.79	1	
35	Br	35	5	80	7	273	2.8	331	266	3.58	1.26	0.07	23.5	1.14	1.24	3.12	1	
36	Kr	36	5	84	8	323	3	121	116	2.16	0.39	0.08	32.2	1.12	2.16	2.6	3	
37	Rb	37	6	86	1	96	0.8	961	312	18.1	0.55	0.08	55.9	2.16	2.48	1.53	5	
38	Sr	38	6	88	2	131	1	1653	1041	33.8	2.1	0.055	33.7	1.91	2.15	2.6	5	
39	Y	39	6	89	3	152	1.3	3200	1782	93	2.7	0.071	19.4	1.62	1.78	4.47	4	
40	Zr	40	6	91	4	160	1.4	3853	2125	120	4	0.066	14.1	1.45	1.6	6.49	3	
41	Nb	41	6	93	5	156	1.6	3573	2741	125	6.4	0.065	10.8	1.34	1.46	8.4	2	
42	Mo	42	6	96	6	166	1.8	5833	2883	128	6.6	0.061	9.4	1.3	1.39	10.2	1	
43	Tc	43	6	98	7	167	1.9	5273	2413	120	5.5	0.06	9	1.27	1.36	11.5	1	
44	Ru	44	6	1	1.1	8	173	2.2	5173	2773	148	6.1	0.057	8.3	1.25	1.34	12.2	2
45	Rh	45	6	1	2.9	8	178	2.2	4773	2239	127	5.2	0.059	8.3	1.25	1.34	12.4	3
46	Pd	46	6	1	6.4	8	192	2.2	4253	1825	90	4	0.058	8.9	1.28	1.37	12	4
47	Ag	47	7	1	7.9	1	175	1.9	2483	1234	60.7	2.7	0.056	10.3	1.34	1.44	10.5	3
48	Cd	48	7	1	12	2	207	1.7	1038	594	23.9	1.46	0.055	13.1	1.48	1.54	8.65	4
49	In	49	7	1	15	3	133	1.7	2273	429	53.7	0.78	0.057	15.7	1.44	1.66	7.31	3
50	Sn	50	7	1	19	4	169	1.8	2543	505	70	1.72	0.054	16.3	1.41	1.62	7.3	3
51	Sb	51	7	1	22	5	199	1.9	1653	904	46.6	4.74	0.049	18.4	1.4	1.59	6.62	2
52	Te	52	7	1	28	6	208	2.1	1263	723	11.9	4.28	0.047	20.5	1.36	1.6	6.24	2
53	I	53	7	1	27	7	241	2.5	456	387	5.2	1.87	0.052	25.7	1.33	1.44	4.94	1
54	Xe	54	7	1	31	8	280	3	165	161	3.02	0.55	0.05	42.9	1.31	2.27	3.06	3

Principal Components Loading Plot

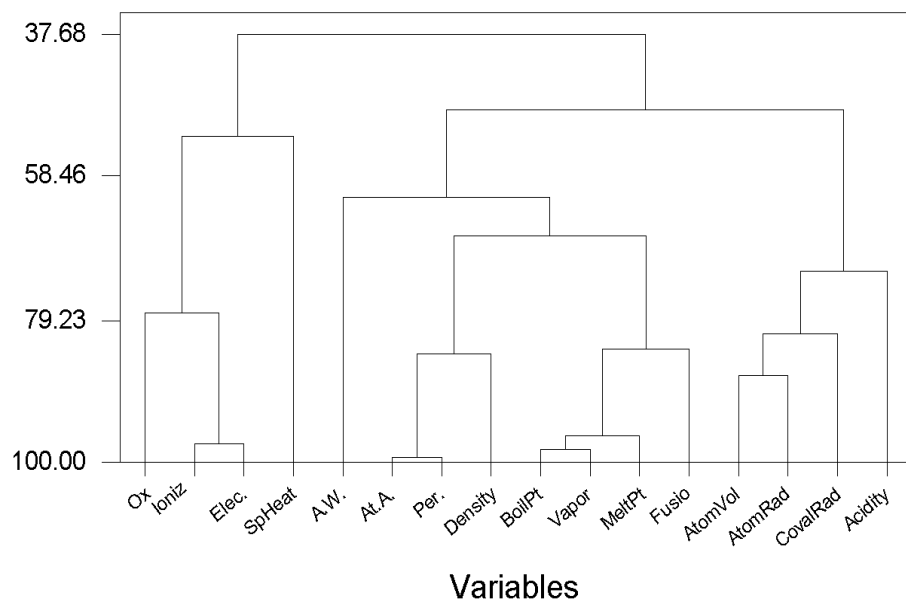


Principal Components Score Plot



Znaky, proměnné
Average, Standard

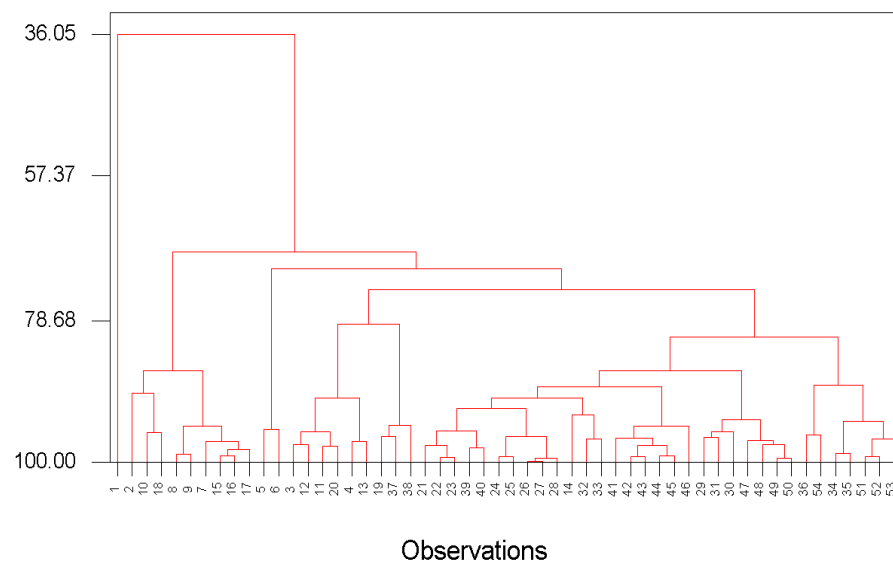
Similarity



Objekty

Similarity

Average, Squared Euclidean, Standard



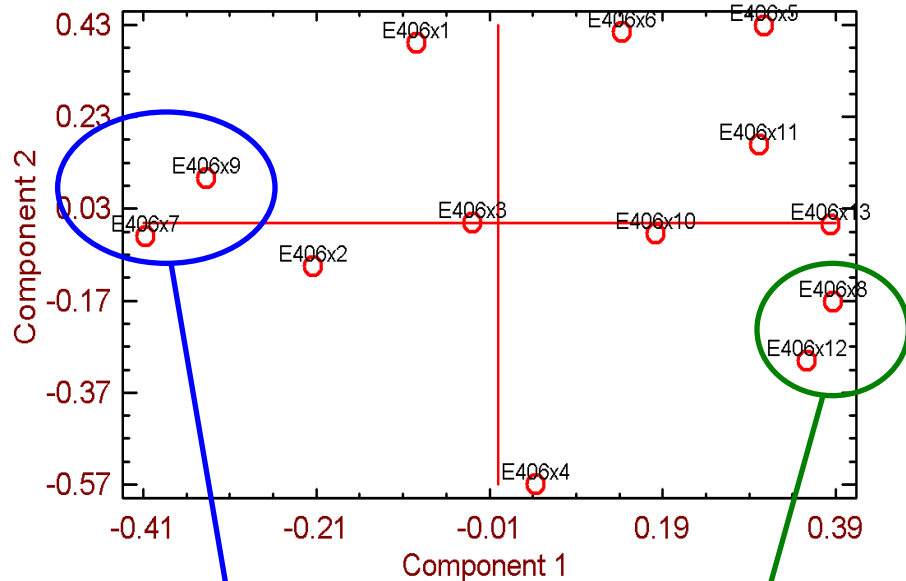
Úloha 6. Klasifikace vlastností rozličných druhů kávy (Kompendium E406)

U 43 vzorků kávy ze 30 zemí byly změřeny chemické a fyzikální vlastnosti. Nalezněte shluky podobných vlastností a shluky podobných prvků.

Data: 13 proměnných (sloupce): **i** index kávy, **j** je **původ kávy**, **x1** obsah vody, **x2** hmotnost zrn, **x3** extrakt, **x4** pH, **x5** volná acidita, **x6** obsah minerálů, **x7** tuky, **x8** kofein, **x9** trinonelin, **x10** kyselina chlorogeniková, **x11** kyselina neochlorogeniková, **x12** kyseliny isochlorogeniková, **x13** suma kyselin chlorogenikových.

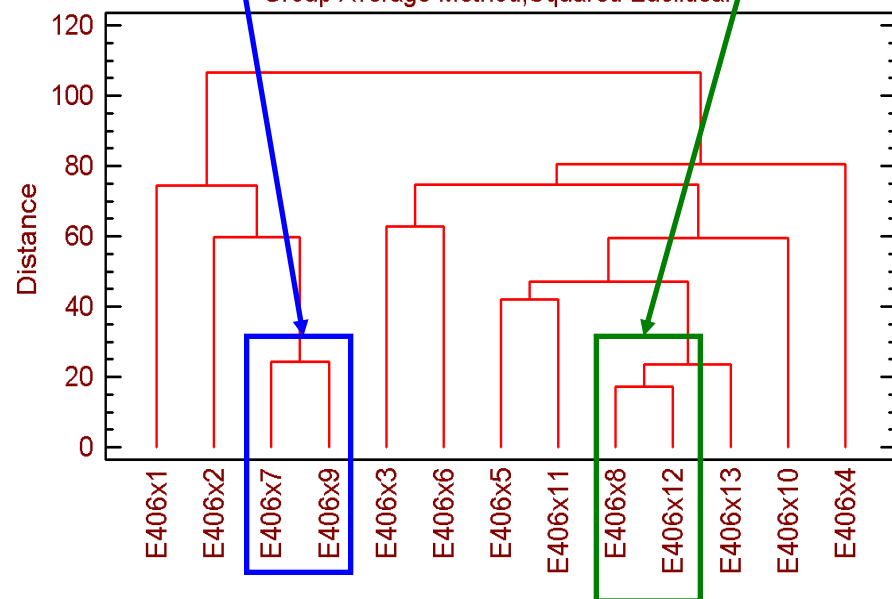
i	ii	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13
1	Mexico 1	8.9	156.6	33.5	5.8	32.7	3.8	15.2	1.1	1	5.4	0.4	0.8	6.6
2	Mexico 2	7.4	157.3	32.1	5.8	30.8	3.7	15	1.3	1	5.1	0.3	1	6.4
3	Guatemala	9.7	152.9	33.1	5.3	36.7	4.2	16.1	1.2	1	5.9	0.2	0.8	6.9
4	Honduras	10.4	174	31.5	5.6	34.2	3.9	15.8	1.1	0.9	5.9	0.4	0.6	6.8
5	Salvador 1	10.5	145.1	35.2	5.8	31.8	4.1	15.2	1.1	1	5.1	0.5	0.7	6.3
6	Salvador 2	10	156.4	34.5	5.8	32.6	3.9	15.4	1.2	0.8	5.3	0.4	0.7	6.4
7	Salvador 3	8.2	155.2	32.4	5.6	29.7	3.8	15.6	1.3	1.2	4.8	0.3	0.7	5.9
8	Nicaragua 1	9.2	167.8	30.6	5.9	28.9	3.8	15.1	1.3	1	5	0.3	0.7	5.9
9	Nicaragua 2	9.3	165.4	35.3	5.8	32.6	4.2	14.3	1.2	1	5.5	0.4	0.8	6.7
10	Costa Rica 1	7.1	180.3	33	5.8	29.3	4	15.1	1.3	1	5.1	0.3	0.7	6.1
11	Costa Rica 2	7.6	153.2	36	5.9	30.5	3.9	16.8	1.4	1.1	5.3	0.3	0.7	6.3
12	Costa Rica 3	7.3	159.6	35	5.8	29.9	3.7	16.5	1.2	1.2	5.5	0.3	0.7	6.5
13	Panama	9.3	161.8	32.4	5.8	31	3.7	15.5	1.3	1.2	5.6	0.3	0.6	6.6
14	Haiti	8.3	160.8	35.7	5.9	30	4.4	13	1.3	1	6.1	0.6	0.8	7.5
15	Dominica	11.6	174.8	32.5	5.4	35.2	3.7	14.5	1	1	5.7	0.3	0.5	6.5
16	Venezuela 1	9.7	169.1	34	5.8	31.6	4	15.7	1.3	1.3	5.1	0.3	0.3	6.2
17	Venezuela 2	10.6	163.7	35	5.8	35	3.8	15.8	1.2	1.1	6.1	0.3	0.9	7.3
18	Columbia 1	12	178.8	32.9	5.3	36.2	4.4	15.6	1.3	1	5.6	0.4	0.7	6.7
19	Columbia 2	10.6	169.1	33	5.3	37.5	4.4	15.1	1.2	1	6.1	0.1	0.6	6.9
20	Ecuador	11.6	148.5	34.6	5.3	39.4	4.2	14.6	1	1.1	5.7	0.5	0.4	6.6
21	Peru	10.1	153.7	34.5	6	28.4	3.7	15.9	1.3	1.1	6.1	0.4	0.8	7.3
22	Brasil 1	10.7	134.5	29.8	5.4	34.1	3.7	15.8	1.2	0.9	5.4	0.4	0.6	6.4
23	Brasil 2	9.7	160.7	33.8	5.3	37.2	4.2	15.2	1.1	0.9	5.4	0.3	0.5	6.2
24	Brasil 3	10.8	133.2	35	5.2	34.7	4.5	15.1	1.2	1.4	5	0.5	0.5	6
25	Brasil 4	11.1	131.7	29.8	5.4	33	4.1	15.8	1.1	1.2	5.1	0.5	0.5	6
26	Brasil 5	10.1	121.6	33.6	5.4	34.7	3.5	15.4	1.1	0.9	5.5	0.4	0.6	6.5
27	Cotedivoir	8	141.8	33.7	5.8	41.9	4.2	11	2	0.5	6.4	0.6	1.5	8.5
28	Togo	9	144.6	29.9	5.6	38	3.9	7.5	1.9	0.3	5.4	0.8	0.9	7.1
29	Cameroon	10.3	119.2	35.5	6.1	41.7	4.1	9.8	1.8	0.8	6	0.5	1.1	7.6
30	Congo	10	143.2	31.7	6.1	29.3	4.1	17	1.2	0.6	5.4	0.3	0.7	6.4
31	Angola 1	9.2	150.4	31.5	5.7	36.4	4.2	8.5	1.9	0.6	5.9	0.6	1.4	7.9
32	Angola 2	9.6	136.6	33.9	5.6	38.2	4	7.2	2.2	0.5	6.2	0.4	1.6	8.3
33	Angola 3	9.5	136.5	32	5.8	31.2	3.8	14.6	1.3	1	5.2	0.4	0.8	6.4
34	Ethiopie	9.3	124.2	35.6	5.8	31.8	3.8	15.7	0.9	0.9	5.5	0.2	0.8	6.5
35	Uganda 1	10.5	132.9	36.2	5.4	36.7	4	15.6	1	1	5.9	0.4	0.6	6.9
36	Uganda 2	10.7	181.2	33.1	5.8	30.7	3.9	15.8	1.3	1.1	5.3	0.3	0.6	6.2
37	Kenya	10.5	159.1	30.3	5.6	31.5	3.7	15.2	1.3	0.9	5.1	0.3	0.7	6
38	Tanganika	9.9	169.4	29	5.6	30.2	3.7	16.5	1.3	0.9	5	0.2	0.7	5.9
39	Madagascar	5	152	30.6	5.3	40.5	3.9	9.6	1.6	0.7	5.3	0.6	0.8	6.7
40	India	11.5	156.8	30.8	5.5	37.5	3.9	14.3	1.2	1	5.8	0.4	0.4	6.6
41	Sumatra	8.4	110.8	31.6	5.7	43.4	4.5	10.1	1.7	0.8	6.3	0.7	0.9	7.9
42	Java	5.6	163.1	34.5	5.5	33.3	4	16	1.2	1.1	5.1	0.3	0.8	6.3
43	Hawai	9.7	191.2	35.1	5.6	34.6	4.2	14.2	1.1	0.9	0.7	0.5	0.3	6.5

Plot of Component Weights

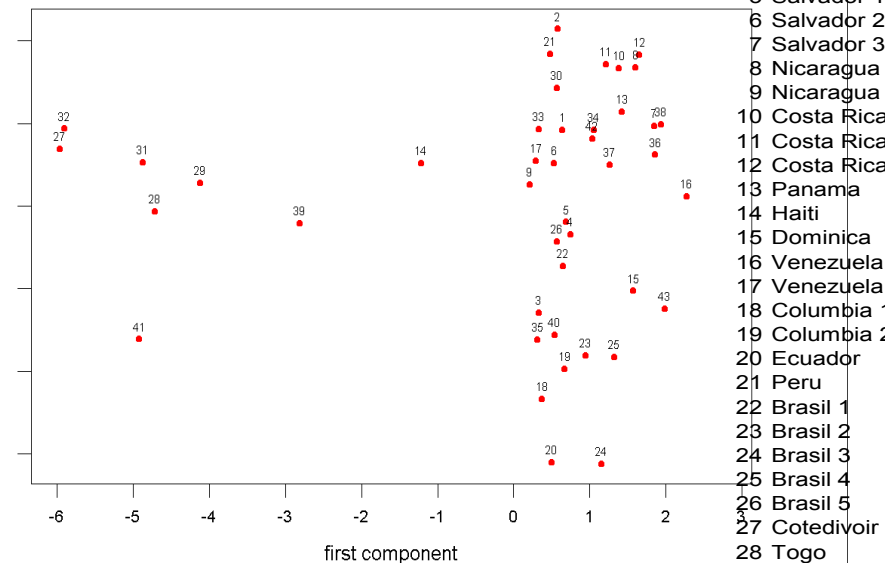


Dendrogram

Group Average Method, Squared Euclidean

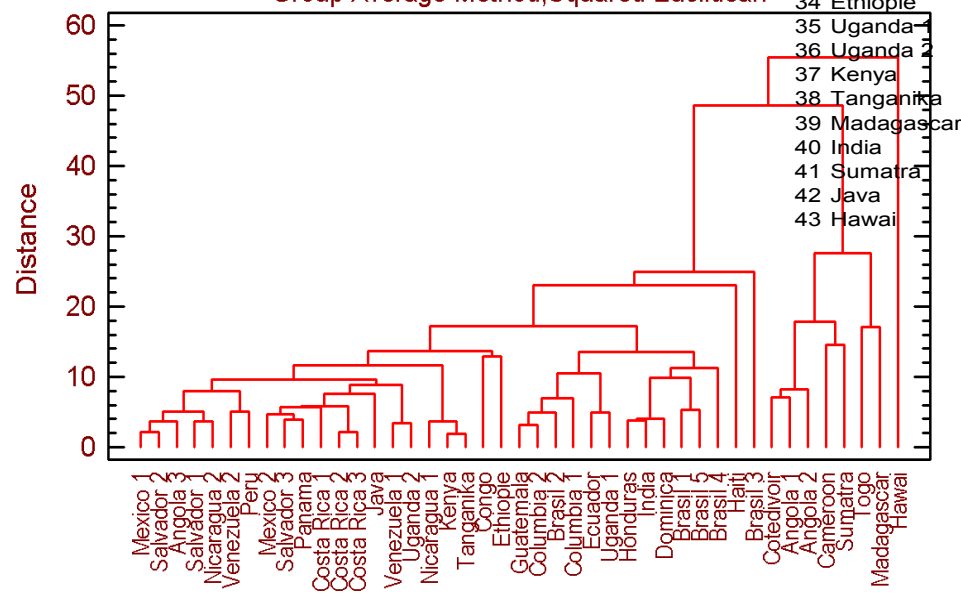


Principal Components Score Plot

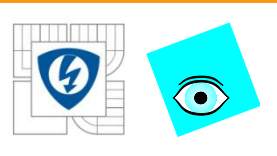


Dendrogram

Group Average Method, Squared Euclidean



- 1 Mexico 1
- 2 Mexico 2
- 3 Guatemala
- 4 Honduras
- 5 Salvador 1
- 6 Salvador 2
- 7 Salvador 3
- 8 Nicaragua 1
- 9 Nicaragua 2
- 10 Costa Rica 1
- 11 Costa Rica 2
- 12 Costa Rica 3
- 13 Panama
- 14 Haiti
- 15 Dominica
- 16 Venezuela 1
- 17 Venezuela 2
- 18 Columbia 1
- 19 Columbia 2
- 20 Ecuador
- 21 Peru
- 22 Brasil 1
- 23 Brasil 2
- 24 Brasil 3
- 25 Brasil 4
- 26 Brasil 5
- 27 Cotedivoir
- 28 Togo
- 29 Cameroon
- 30 Congo
- 31 Angola 1
- 32 Angola 2
- 33 Angola 3
- 34 Ethiopie
- 35 Uganda 1
- 36 Uganda 2
- 37 Kenya
- 38 Tanganika
- 39 Madagascar
- 40 India
- 41 Sumatra
- 42 Java
- 43 Hawaii



Postup analýzy vícerozměrných dat

1. Standardizace: analýze vždy předchází standardizace čili škálování proměnných.

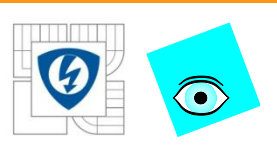
2. Odhady parametrů polohy, rozptýlení, tvaru a intenzita vztahu mezi proměnnými:

Vyčíslení výběrové střední hodnoty každé proměnné.

Odhad kovarianční matice $\mathbf{S}$ a její normované podoby - korelační matice $\mathbf{R}$.

Odhadu vícerozměrné šikmosti a vícerozměrné špičatosti.

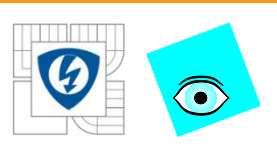
Matice $\mathbf{R}$ obsahuje Pearsonovy párové korelační koeficienty, které se diskutují.



Postup analýzy vícerozměrných dat

3. Exploratorní analýza dat EDA:

- a) Hledání podobnosti objektů vizuálními rozptylovými diagramy typu **casement plot**, **draftsman plot**, dále symbolových a profilových grafů (**hvězdičky**, **sluníčka**, **obličej**e, **křivky**, **stromy**),
- b) Nalezení vybočujících objektů nebo vybočujících proměnných, mnohdy nevhodných k analýze,
- c) Testy předpokladů lineárních vazeb,
- d) Testy předpokladů o datech (normalitu, nekorelovanost, homogenitu).
Ověřování normality založené na vícerozměrné šikmosti a vícerozměrné špičatosti.



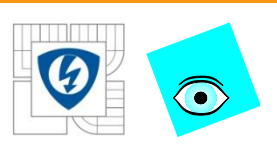
Postup analýzy vícerozměrných dat

4. Určení vhodného počtu latentních proměnných:

- a) Matice S nebo R se rozloží na **vlastní čísla** a **vlastní vektory**.
- b) **Indexový graf úpatí vlastních čísel** (Scree plot): určí vhodný počet latentních proměnných, které ještě dostatečně popisují proměnlivost v datech.
- c) Když se latentní proměnné podaří **pojmenovat** a dát jim i fyzikální, biologický či jiný věcný význam, jedná se o faktory. Jinak jde o hlavní komponenty.

5. Určení struktury v proměnných (PCA a FA):

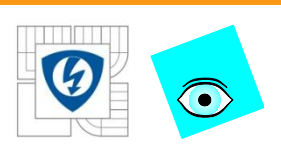
- a) **Graf komponentních vah** (Plot of components weights, loadings): hledání struktury a vzájemných vazeb (korelace) proměnných se provede v grafu
- b) **Rozptylový diagram komponentního skóre** (Scatterplot): hledání struktury v objektech a třídění objektů do shluků.
- c) **Dvojný graf** (Biplot) je přehledným spojením obou předešlých grafů a ukáže interakci objektů a proměnných.



Postup analýzy vícerozměrných dat

6. Určení struktury a vzájemných vazeb v objektech:

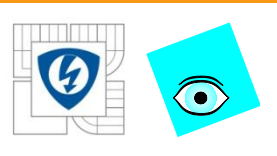
- a) Klasifikační postupy **zařadí analyzovaný objekt** do jednoho již existujícího a předem zadaného shluku.
- b) Neutříděnou skupinu objektů lze **uspořádat do shluků** a výsledek třídění zobrazit dendrogramem v analýze shluků. V hierarchickém postupu je třeba k vytvoření shluků vybrat vzdálenost mezi objekty (Eukleidovskou, Manhattanovskou, Mahalanobisovu) a jednu z nabídnutých metod: průměrovou, centroidní, nejbližšího souseda, nejvzdálenějšího souseda, mediánovou, Wardovu.
- c) Nehierarchické postupy rozdělí objekty do shluků, v nichž jsou předem umístění **typičtí reprezentanti**.



Postup analýzy vícerozměrných dat

7. Vysvětlení souladu nalezené struktury objektů a vzájemných vazeb v dendrogramu a PCA (či FA) grafech:

- a) Vyšetřit a vysvětlit nalezenou **strukturu a vazby jednotlivých proměnných** nalezenou jednak v PCA (či FA) a jednak v dendrogramu podobnosti proměnných analýzou vzniklých shluků.
- b) Vysvětlit **strukturu a vazby klasifikovaných objektů** nalezenou v PCA a v dendrogramu podobnosti objektů.



Cvičení v programu STATISTICA:

51. Zdrojová matice *factor.sta* se načte z databáze EXAMPLES programu STATISTICA.

Znaky: úloha se týká výběru různých automobilů, od každého byly zaznamenány následující znaky (proměnné):

1. Price značí přibližnou cenu vozu v tisících dolarů.
2. Acceler značí akceleraci vozu od 0 do 60 sec.
3. Braking značí brzdění vozu z rychlosti 80mph do naprostého klidu.
4. Handling značí index vůči stavu silnice.
5. Mileage značí spotřebu benzínu v mílích a na galon.

Standardizace: protože je každý znak uveden v jiných jednotkách (sekundy, tisíce dolarů, mpg, atd.), je nutné data předem standardizovat tak, aby každý znak měl průměr 0 a rozptyl 1. Standardizovaná data jsou uložena v souboru *car.sta*.

Cíl analýzy: lze určit taxonomii (systematiku třídění) v této klasifikaci aut? Vytvoří tyto vozy přirozené shluky, které budou smysluplně nadepsány? Provedeme zde rovněž spojovací shlukování a pak hierarchické shlukování?

STATISTICA Cz - [Data: Cars (5s krát 22ř)]

Soubor Úpravy Zobrazit Vložit Formát Statistika Grafy Nástroje Data Okno Nápověda

Arial 10 B I U

Přidat do sešitu Přidat do protokolu

Proměnné Případy

Performance, fuel economy, and approximate price for various automobiles					
	1	2	3	4	5
	PRICE	ACCELERATION	BRAKING	HANDLING	MILEAGE
Acura	-0,521	0,477	-0,007	0,382	2,079
Audi	0,866	0,208	0,319	-0,091	-0,677
BMW	0,496	-0,802	0,192	-0,091	-0,154
Buick	-0,614	1,689	0,933	-0,210	-0,154
Corvette	1,235	-1,811	-0,494	0,973	-0,677
Chrysler	-0,614	0,073	0,427	-0,210	-0,154
Dodge	-0,706	-0,196	0,481	0,145	-0,154
Eagle	-0,614	1,218	-4,199	-0,210	-0,677
Ford	-0,706	-1,512	0,887	0,145	1,734
Honda	-0,429	0			
Isuzu	-0,798	0			
Mazda	0,126	0			
Mercedes	1,051	0			
Mitsub.	-0,614	-1			
Nissan	-0,429	0			
Olds	-0,614	-0			
Pontiac	-0,614	0			
Porsche	3,454	-2			
Saab	0,588	0			
Toyota	-0,059	1			
VW	-0,706	-0			
Volvo	0,219	0			

Otevřít

Oblast hledání: Datasets

Poslední dokumenty
 Plocha
 Dokumenty
 Tento počítač
 Místa v síti

Aircraft
 Allergy
 AnalyzingSICCodes
 Ancova
 Animal Weights
 Ball bearings
 Barotro2
 Barotrop
 Barotropsn
 Baseball
 BatchesForDeployment
 Beverag2
 Beverage
 Bid_prep
 Bleach
 Boston2
 BostonHousing
 Brake pad
 Bulbs
 Car Accidents
Cars
 Cat Clinic
 Cat Clinic Measurements (2000)
 Cat Clinic Patient Info
 Ccd
 Center
 Center2
 Circuits
 College
 Comfort
 Composit
 Computers
 Constr
 Coupons
 Cover
 Cover2
 Crabs
 Cracker promc
 Credit applicat
 CreditSNN
 Crossover
 Cusum
 Cycling
 Damage
 Demand

Název souboru: Cars

Soubory typu: STATISTICA soubory (*.stw;*.sta;*.stg;*.str;*.svl)

Otevřít

Storno

52. Načteme soubor *Cars.sta*.

53. Soubor je předem standardizován.

Soubor Úpravy Zobrazit Vložit Formát Statistika Grafy Nástroje Data Okno Nápověda

Obnovit... Ctrl+R

Analýza skupin

- Základní statistiky/tabulky
- Vícerozměrná regrese
- ANOVA
- Neparametrická statistika
- Prokládání rozdělení
- Pokročilé lineární/nelineární modely
- Vícerozměrné průzkumné techniky**
- Průmyslová statistika & Six Sigma
- Analýza síly testu
- Neuronové sítě
- Vytěžování dat
- QC Data mining & Analýza hlavních příčin
- Text & Document Mining, Web Crawling
- Statistiky bloku dat
- STATISTICA Visual Basic
- Pravděpodobnostní kalkulátor

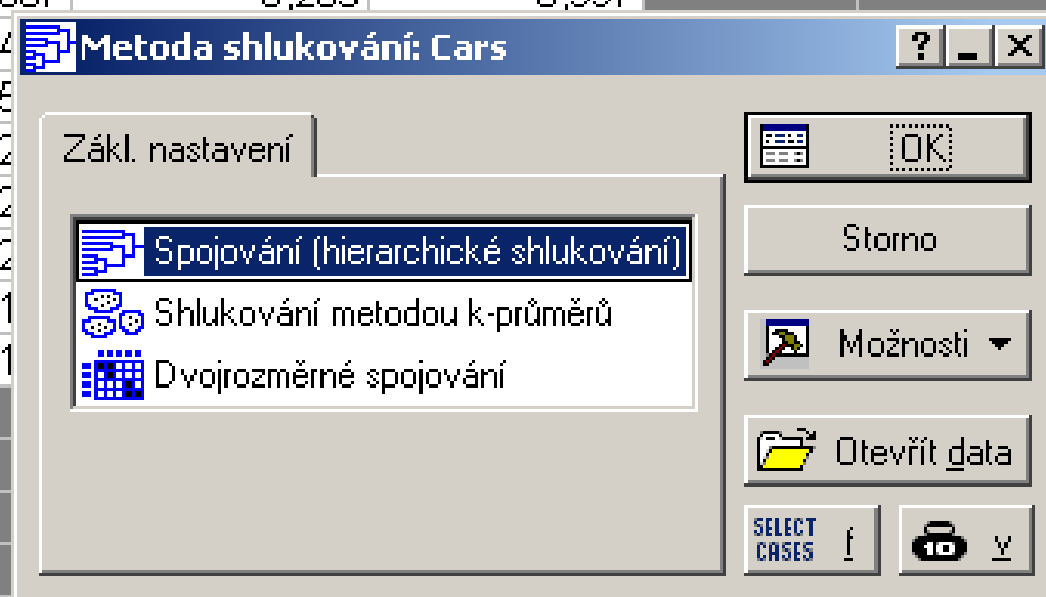
Shluková analýza

- Faktorová analýza
- Hlavní komponenty & klasifikační analýza
- Kanonická analýza
- Analýza spolehlivosti/prvků
- Klasifikační stromy
- Korespondenční analýza
- Vícerozměrné škálování
- Diskriminační analýza
- Modely obecné diskriminační analýzy

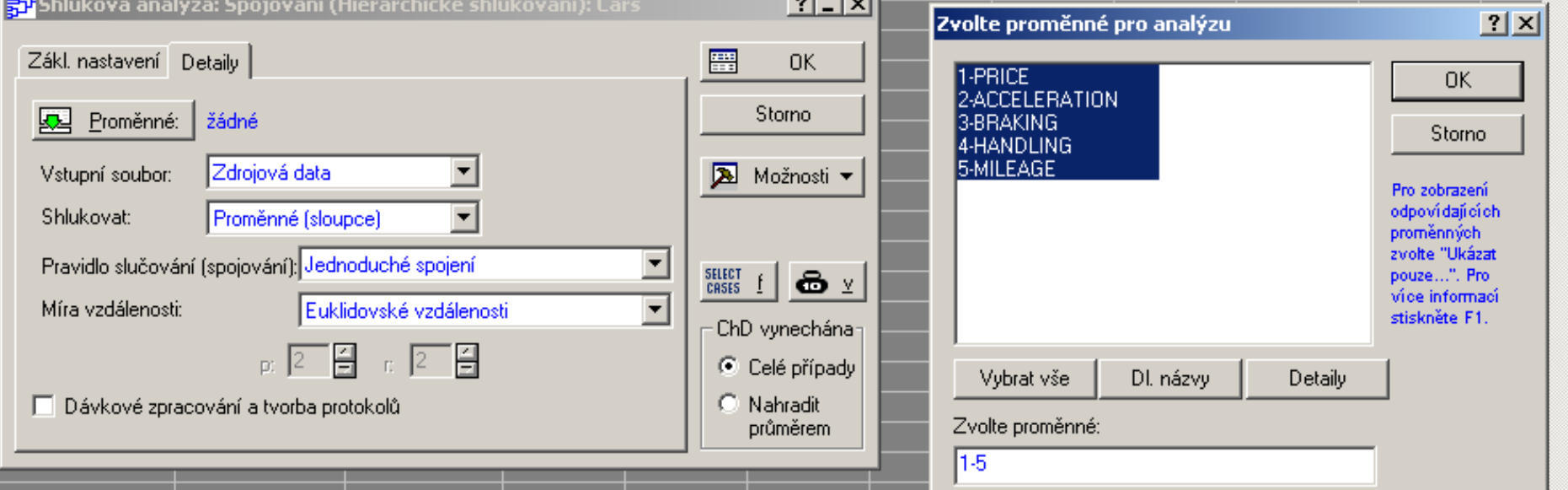
	Performanc	1	2	3	4	5
	PRICE				HANDLING	MILEAGE
Acura	-0				0,382	2,079
Audi	0				-0,091	-0,677
BMW	0				-0,091	-0,154
Buick	-0				-0,210	-0,154
Corvette	1					
Chrysler	-0					
Dodge	-0					
Eagle	-0					
Ford	-0					
Honda	-0					
Isuzu	-0					
Mazda	0					
Mercedes	1					
Mitsub.	-0					
Nissan	-0					
Olds	-0					
Pontiac	-0					
Porsche	3,454	-2,215	-0,296		0,618	-1,026
Saab	0,588	0,679	0,246		0,263	0,021
Toyota	-0,059	1,218	0,228		0,736	-0,851
VW	-0,706	-0,128	0,102		0,382	0,195
Volvo	0,219	0,612	0,138		-0,210	0,369

54. Zavoláme *Shlukovou analýzu* z bloku *Vícerozměrné průzkumné techniky*.

,410	-0,007	0,027	0,369
,410	-0,061	-4,230	1,067
,679	-0,133	0,500	-1,724
,006	0,120	-0,091	-0,154
,003	0,084	0,382	0,718
,073	-0,007	0,263	0,997



55. V okně *Metoda shlukování* zvolíme v *Základním nastavení* jednu ze tří nabídnutých metod, a to techniku *Hierarchického shlukování*.



56. V okně *Shluková analýza* vezmeme záložku *Detaily* v okénku *Proměnné* načteme 5 znaků. Protože načítáme data ve formě **zdrojové matice dat**, musíme to také uvést v okénku *Vstupní soubor* slovem *Zdrojová data*. V okénku *Shlukovat* máme dvě možnosti zadání: *Proměnné (sloupce)*, kde zadáme shlukování znaků do dendrogramu nebo *Případy (řádky)*, kde zadáme shlukování objektů, zde aut. V okénku *Pravidlo slučování* zadáme užitou shlukovací metodu. Začneme s metodou *Jednoduché spojení*, což je vlastně metoda nejbližšího souseda. V okénku *Míra vzdálenosti* zadáme typ vyčíslované vzdálenosti, tj. nejčastěji užívané *Euklidovské vzdálenosti*, počítané dle Pythagorovy věty. Když bychom zaškrtnli *Dávkové zpracování a tvorba protokolu*, vypracovaly by se automaticky tři okna výsledků. Zatím necháme toto okénko volné a budeme zadávat analýzy ručně. Jsou-li v datech díry, ve kterých nějaké hodnoty znaků chybí, pak tento objekt čili řádek vynecháme. To se zadá zaškrtnutím *Celé případy* a pak OK.

Počet proměnných: 5
Počet případů: 22
Spojování proměnných
Chybějící data: odstr. případově
Pravidlo slučování (spoj.): Jednoduché spojení
Metrika vzdálenosti: Euklid. vzdálenosti (nestandardiz.)

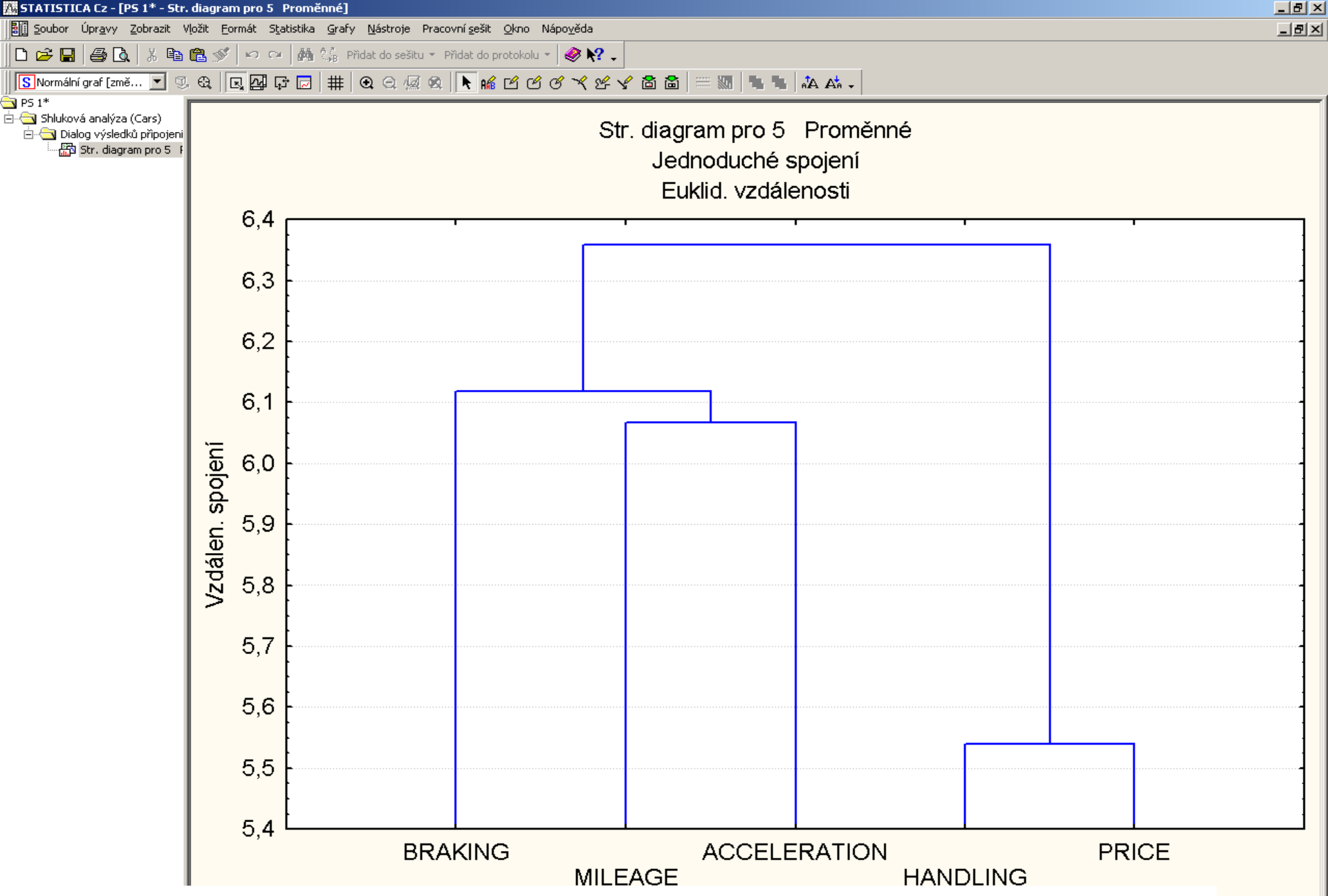
Zákl. výsledky Details

Horizontální graf hierarch. stromu
Vertikální "třásňový" graf
☒ Pravoúhlé větve
☐ Standardizovat měřítko stromu (*100)
Rozvrh shlukování
Graf rozvrhu shlukování

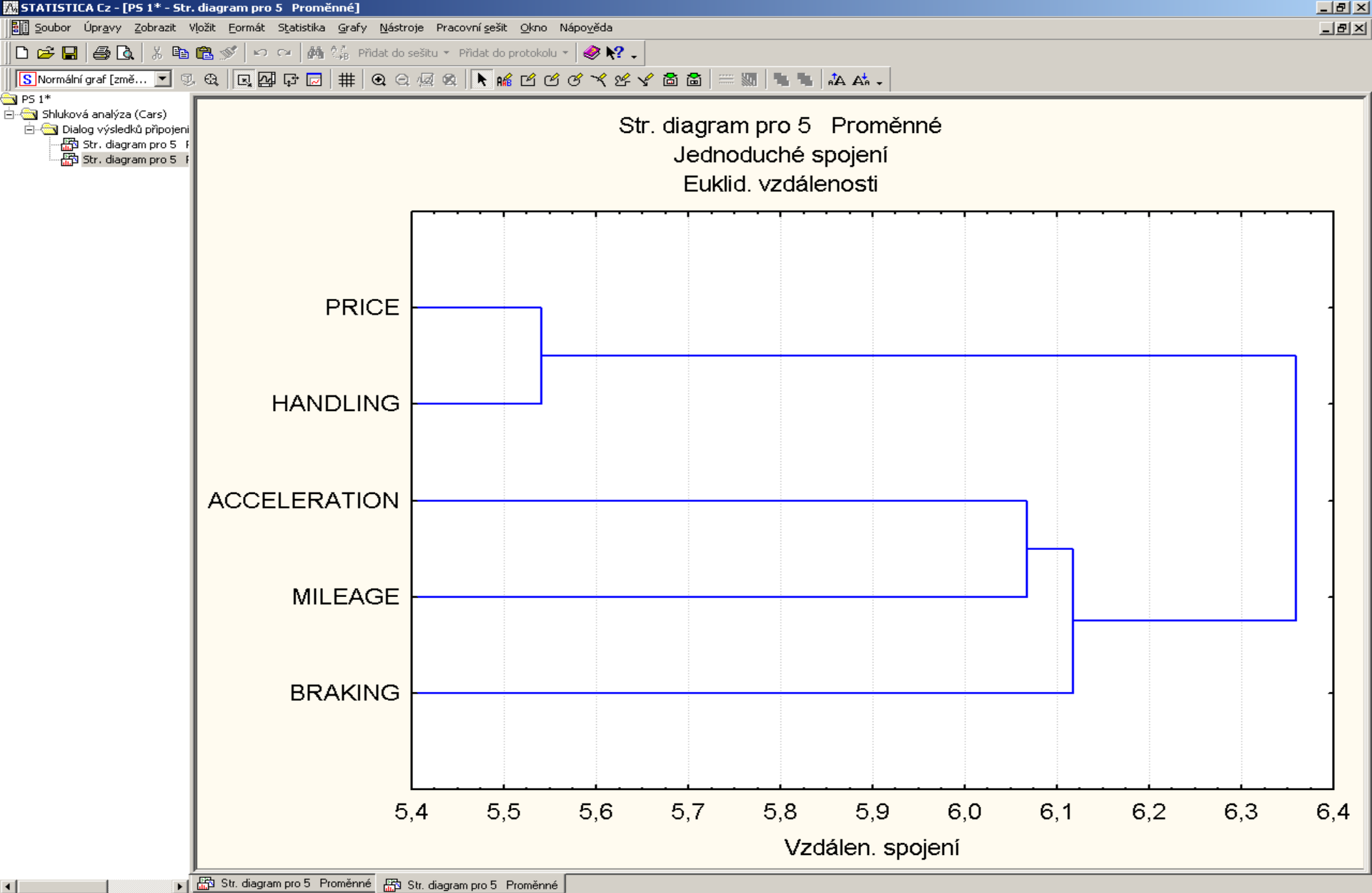
Matice vzdáleností
Popisné statistiky
Matice

Souhrn
Storno
Možnosti ▼

57. *Výsledky spojování:* v horní polovině okna je rekapitulace podmínek výpočtu. V dolní polovině jsou pak dvě záložky: *Zákl. Nastavení* a *Details*. V *Detailech* klikneme na *Vertikální třásňový graf*, když jsme předtím zaškrtnuli okénko *Pravoúhlé větve*. Zkusme si ale také zopakovat toto bez tohoto zaškrtnutí. Zde budeme nyní postupně klikat na všechna okénka a záložky a seznamovat se s výsledky.



58. Dostaneme dendrogram proměnných (znaků) čili vývojový strom shluků ve vertikálním tvaru.



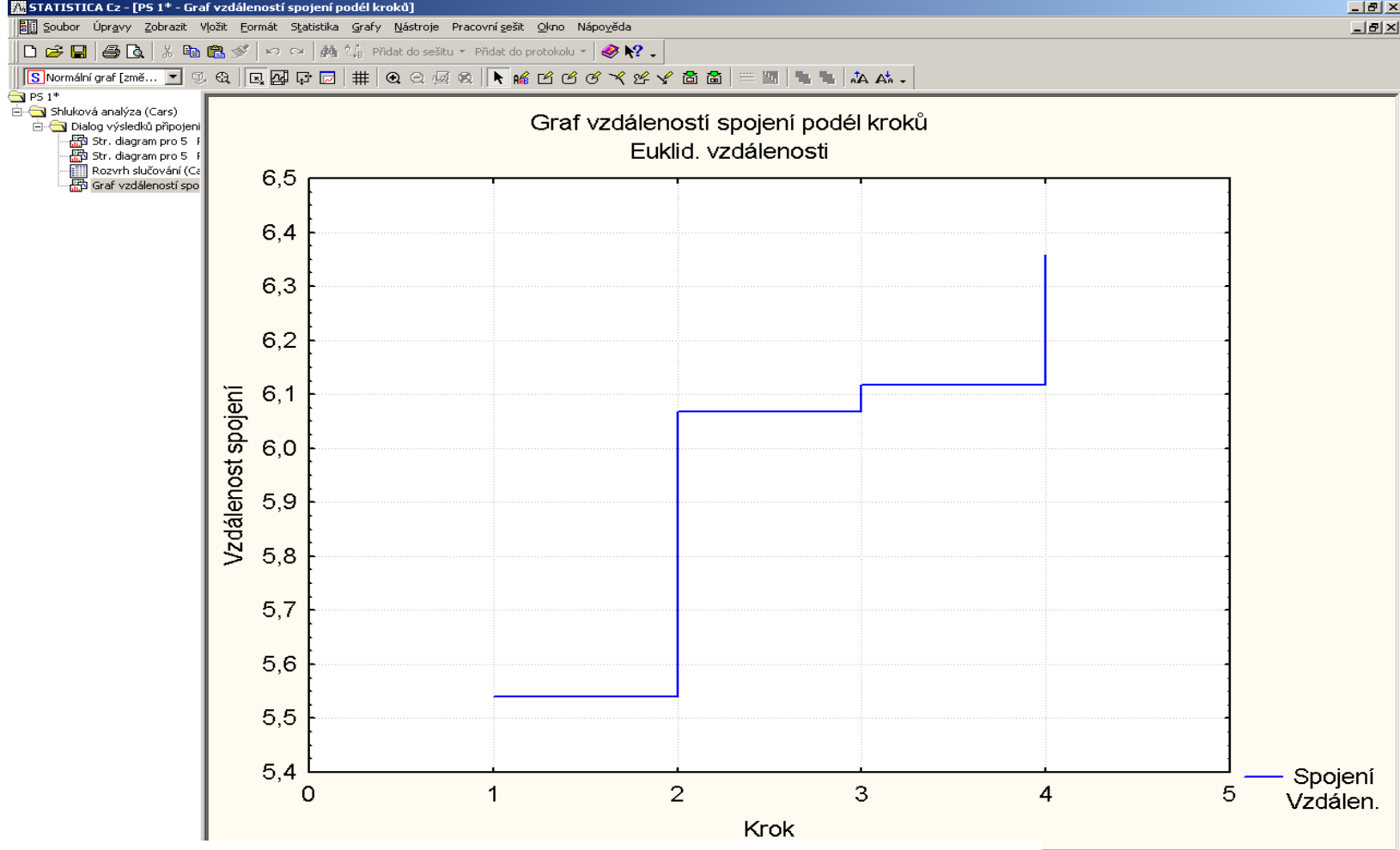
59. Když klikneme na okénko *Horizontální graf hierarchického stromu* dostaneme tento dendrogram (strom) v tomto horizontálním uspořádání.

PS 1*

- Shluková analýza (Cars)
 - Dialog výsledků připojení
 - Str. diagram pro 5 f
 - Str. diagram pro 5 f
 - Rozvrh slučování (Cars)

Rozvrh slučování (Cars)							
Jednoduché spojení							
Euklid. vzdálenosti							
spojení vzdálen.	Obj. č. 1	Obj. č. 2	Obj. č. 3	Obj. č. 4	Obj. č. 5		
5,540081	PRICE	HANDLING					
6,067299	ACCELERATION	MILEAGE					
6,118024	ACCELERATION	MILEAGE	BRAKING				
6,358982	PRICE	HANDLING	ACCELERATION	MILEAGE	BRAKING		

60. *Rozvrh slučování:* když v okénku *Výsledky spojování* klikneme na okénko *Rozvrh slučování*, dostaneme tuto tabulku negrafické podoby postupného vytvoření dendrogramu.



61. *Graf vzdáleností spojení podél kroků*: když v okénku *Výsledky spojování* klikneme na *Graf rozvrhu slučování*, objeví se tento graf. Jeví se užitečným v operativním zastavení tvorby dendrogramu. V horizontálním dendrogramu platí pravidlo, čím více doprava, tím vzdálenější shluky jsou připojovány. Když ale v tomto grafu rozvrhu slučování najdeme plateau, znamená to, že řada shluků byla vytvořena v podstatě při stejné spojovací vzdálenosti. Tato vzdálenost se pak jeví optimální k ukončení tvorby shluků při rozhodování o nejvhodnějším počtu shluků dendrogramu.

PS 1*

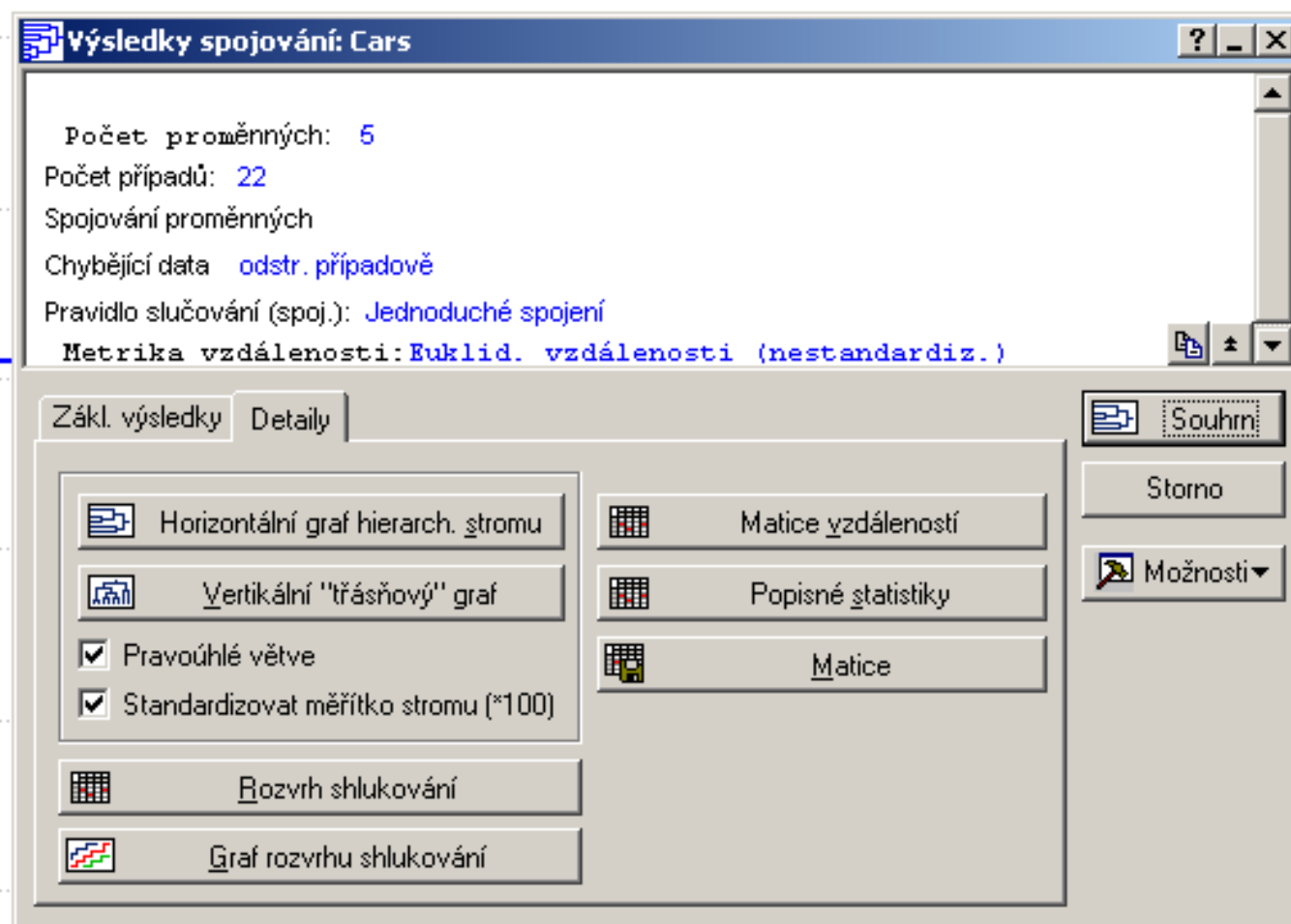
- Shluková analýza (Cars)
 - Dialog výsledků připojení
 - Str. diagram pro 5 F
 - Str. diagram pro 5 F
 - Rozvrh slučování (Ca
 - Graf vzdáleností spo
 - Euklid. vzdálenosti (C

	Euklid. vzdálenosti (Cars)				
Proměnná	PRICE	ACCELERATION	BRAKING	HANDLING	MILEAGE
PRICE	0,00	7,90	6,62	5,54	7,60
ACCELERATION	7,90	0,00	6,97	7,18	6,07
BRAKING	6,62	6,97	0,00	6,36	6,12
HANDLING	5,54	7,18	6,36	0,00	7,22
MILEAGE	7,60	6,07	6,12	7,22	0,00

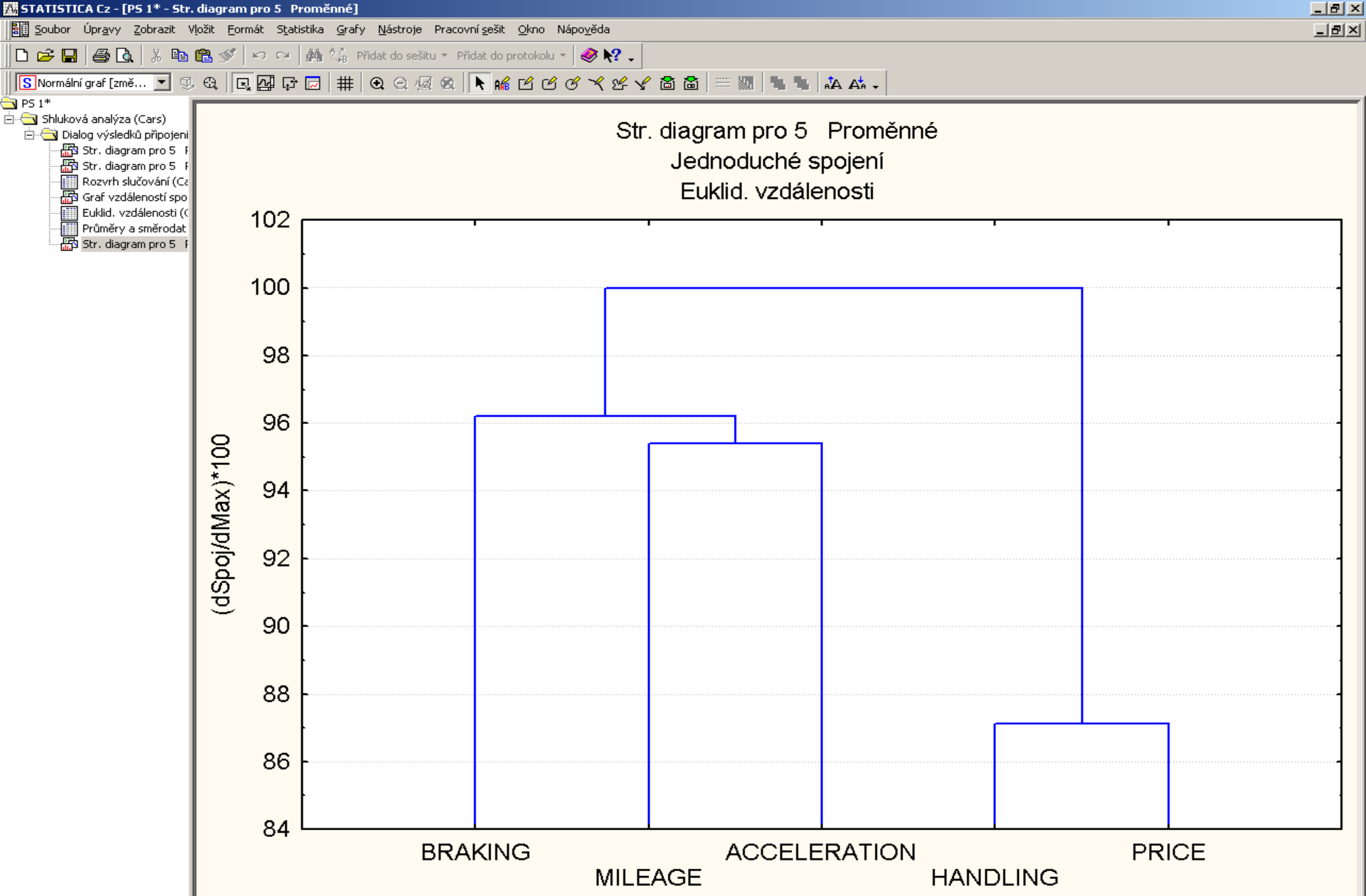
62. *Euklidovské vzdálenosti:* když v okénku *Výsledky spojování* klikneme na *Matice vzdálenosti*, otevře se nám tato tabulka *Euklidovské vzdálenosti*.

	Cars									
	1 PRICE	2 ACCELERATION	3 BRAKING	4 HANDLING	5 MILEAGE					
PRICE	0,00000	7,89893	6,62279	5,54008	7,59799					
ACCELEF	7,89893	0,00000	6,96909	7,18369	6,06730					
BRAKING	6,62279	6,96909	0,00000	6,35898	6,11802					
HANDLIN	5,54008	7,18369	6,35898	0,00000	7,22414					
MILEAGE	7,59799	6,06730	6,11802	7,22414	0,00000					
Průměry	0,00000	0,00000	-0,00000	-0,00000	0,00000					
Sm.Odch	1,00000	1,00000	1,00000	1,00000	1,00000					
Poč.přip	22,00000									
Matice	3,00000									

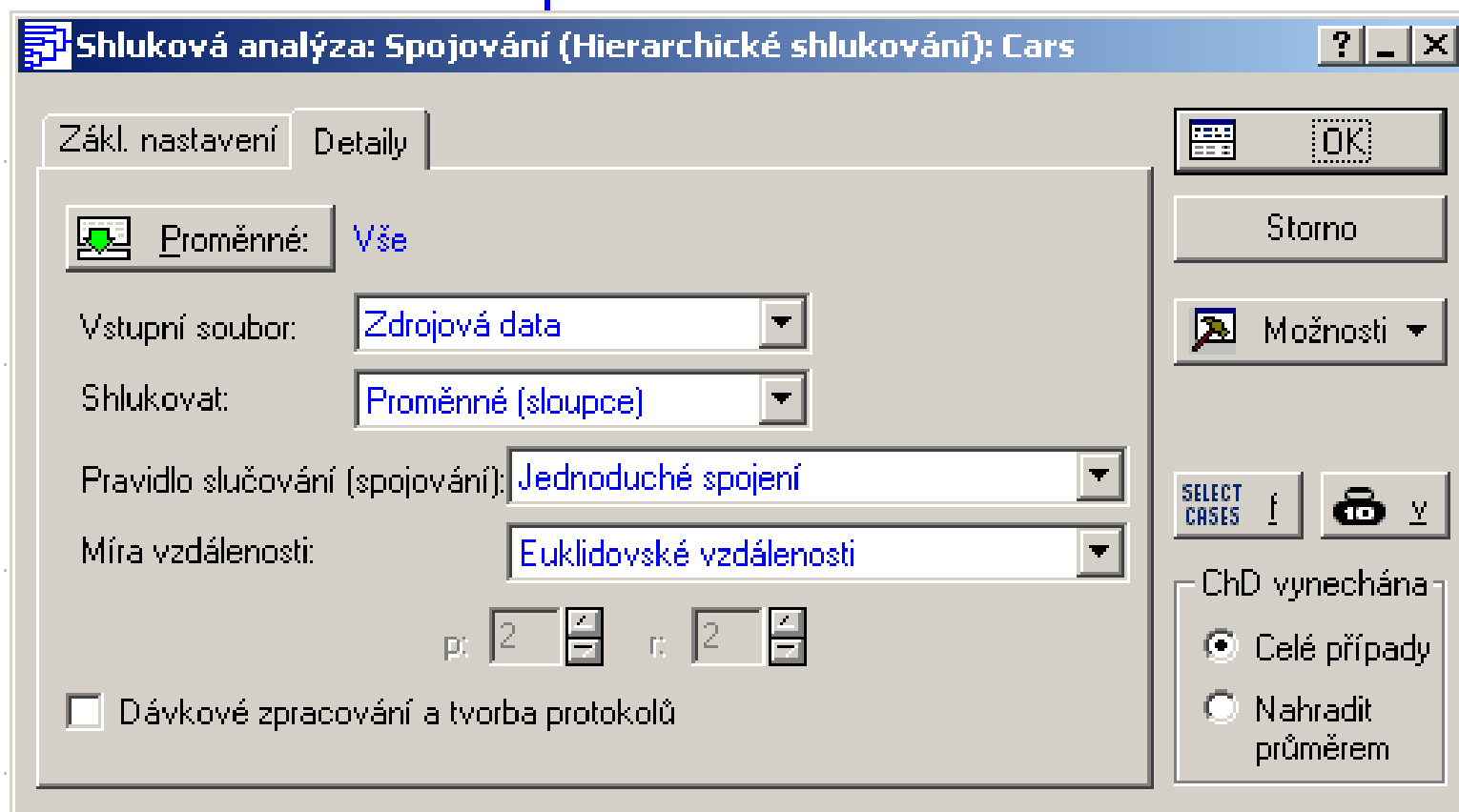
63. *Cars*: když v okénku *Výsledky...* klikneme na záložku *Matice*, dostaneme tuto tabulku, po uložení použitelnou i v dalších výpočtech vícerozměrné statistické analýzy. Je to matice Euklidovských vzdáleností užívaných znaků všech aut.



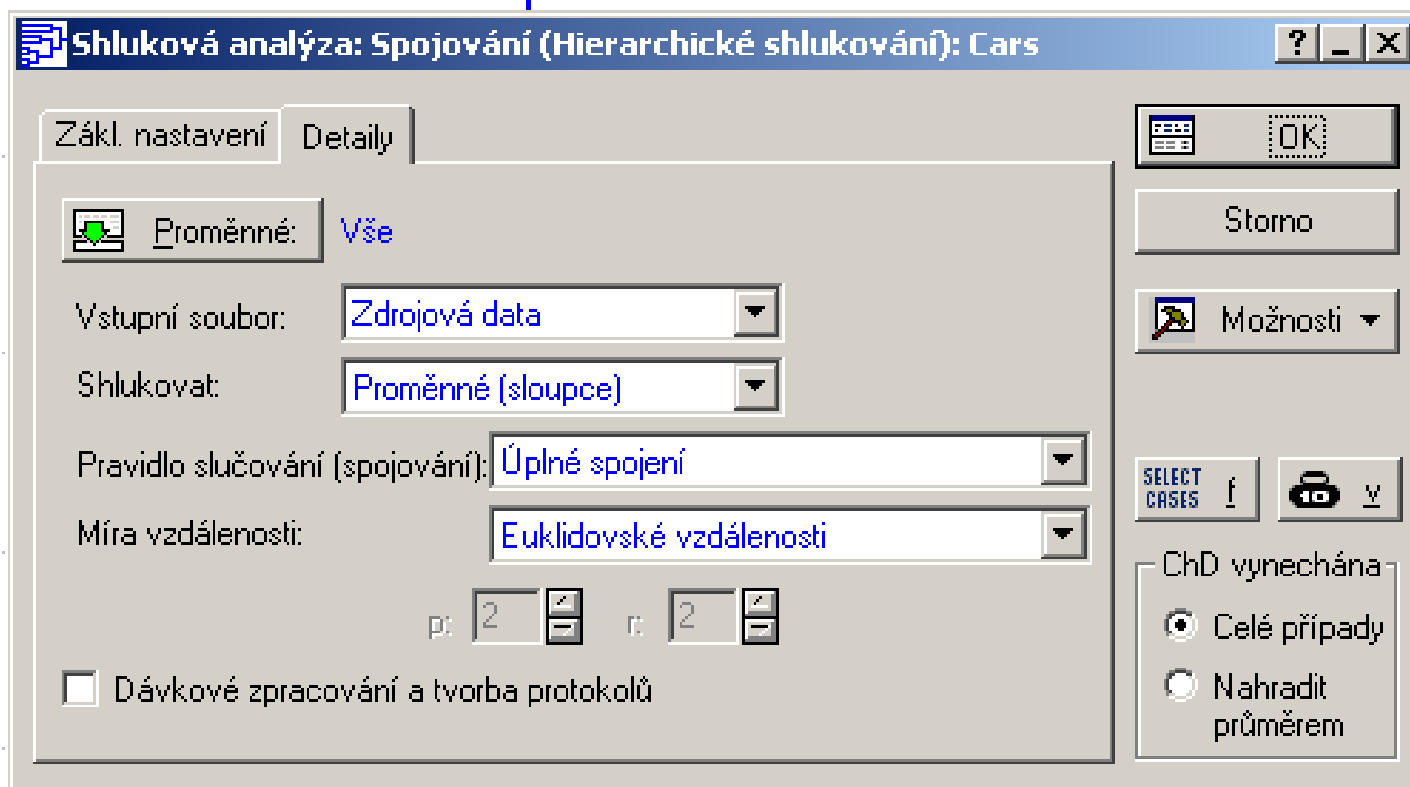
64. V okénku *Standardizovat měřítko stromu* zaškrtneme záložku a pak klikneme na *Vertikální trásňový graf*.



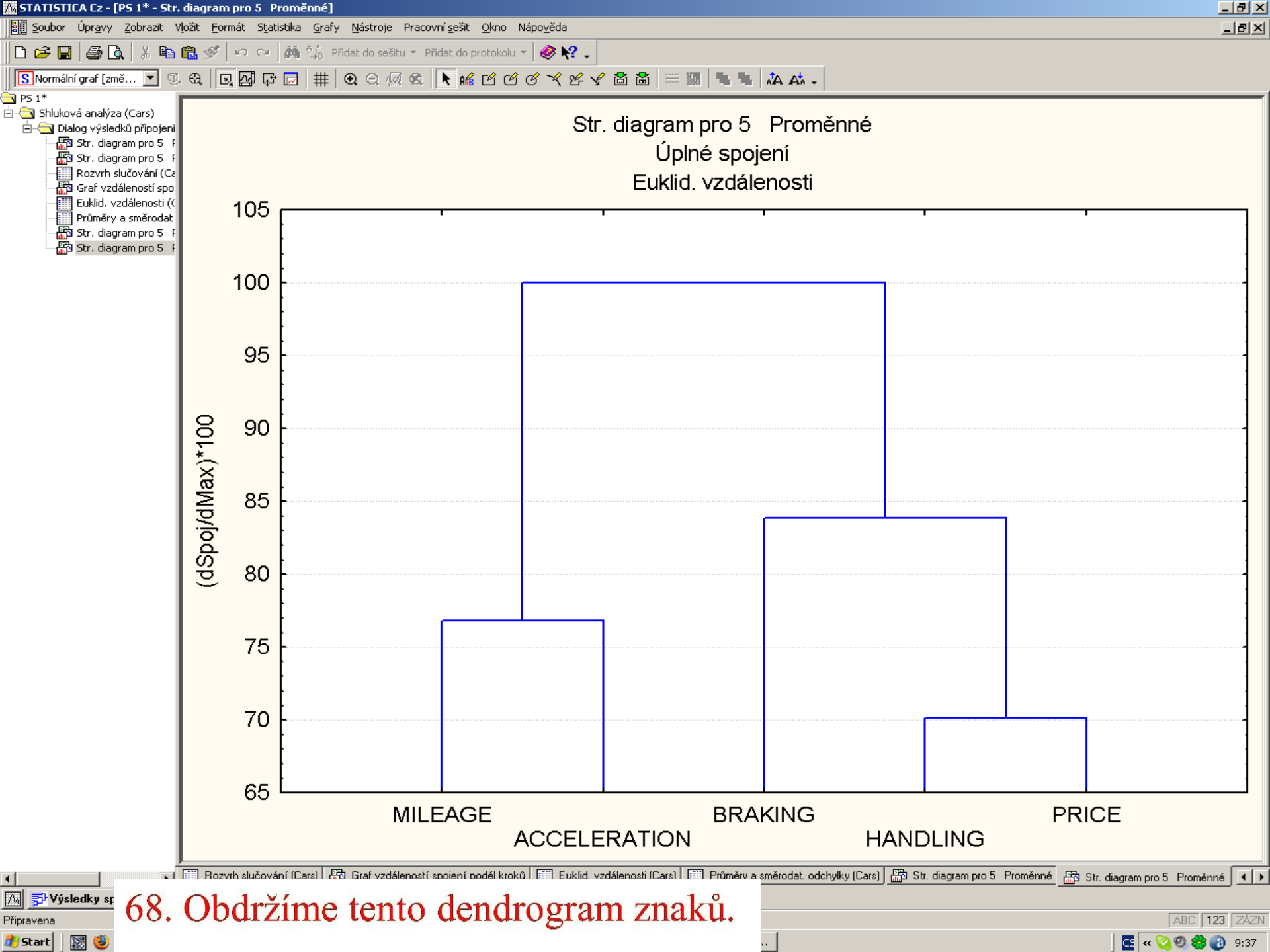
65. Otevře se nám dendrogram v relativních čili normovaných vzdálenostech v procentech na y ose.

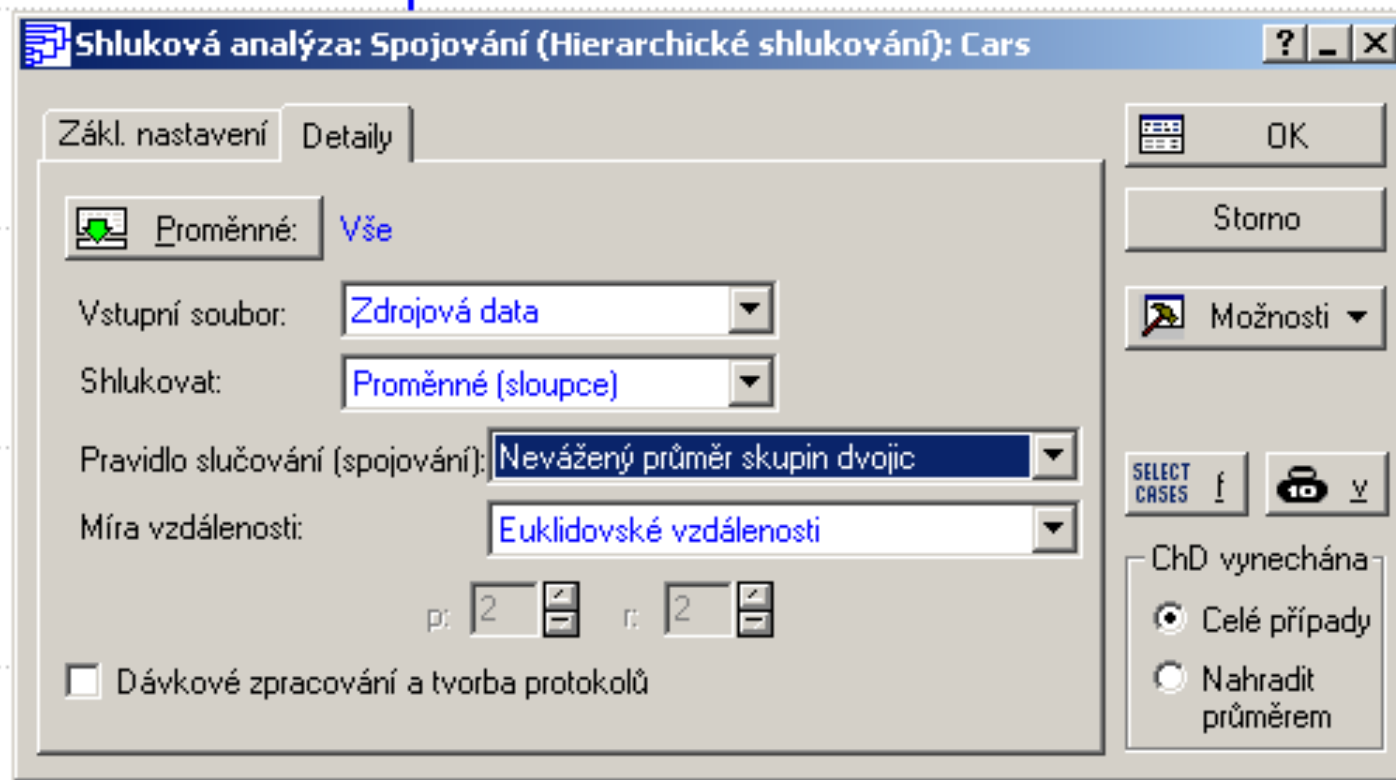


66. Navrátíme se do tohoto výchozího okénka zadání tvorby dendrogramu.

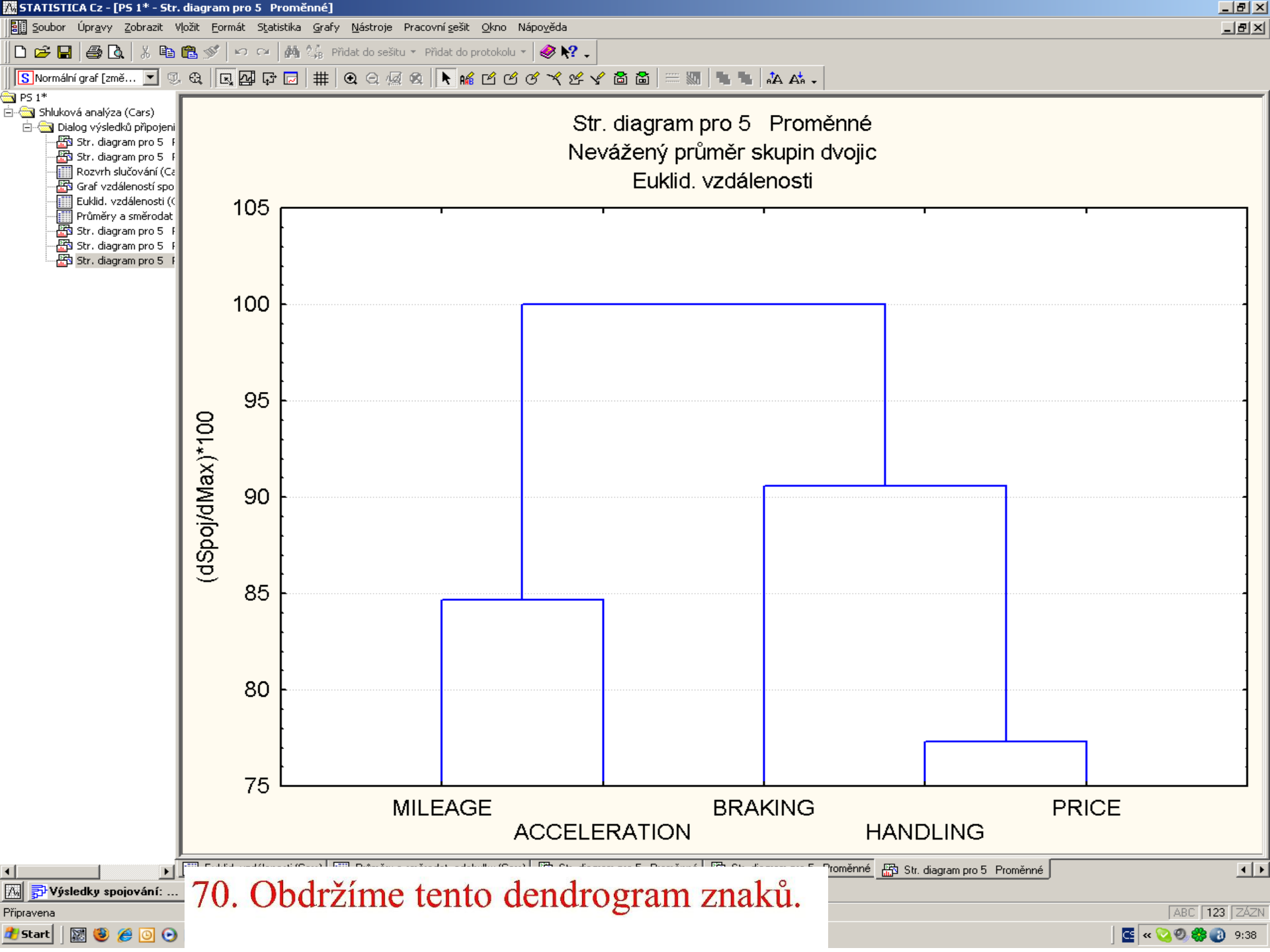


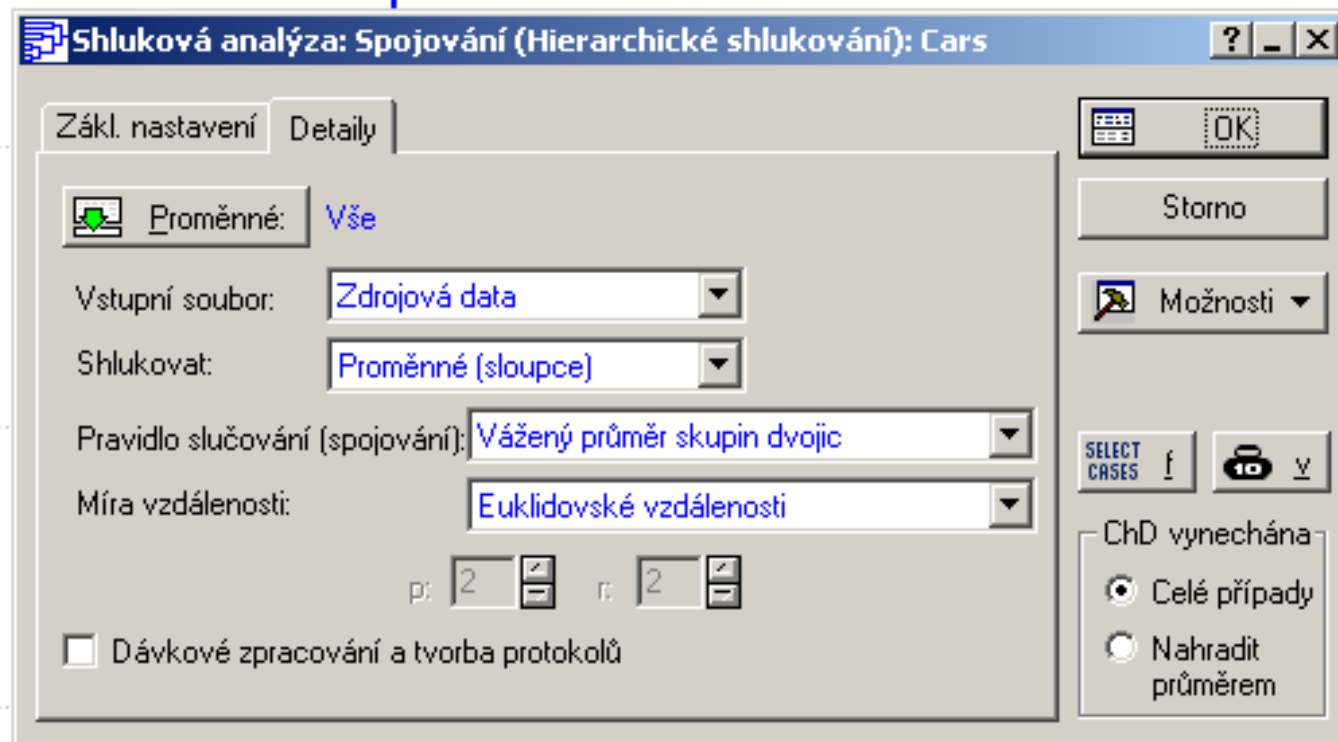
67. Zadáme jinou shlukovací metodu *Úplné spojení*, což je metoda nejvzdálenějšího souseda a OK.



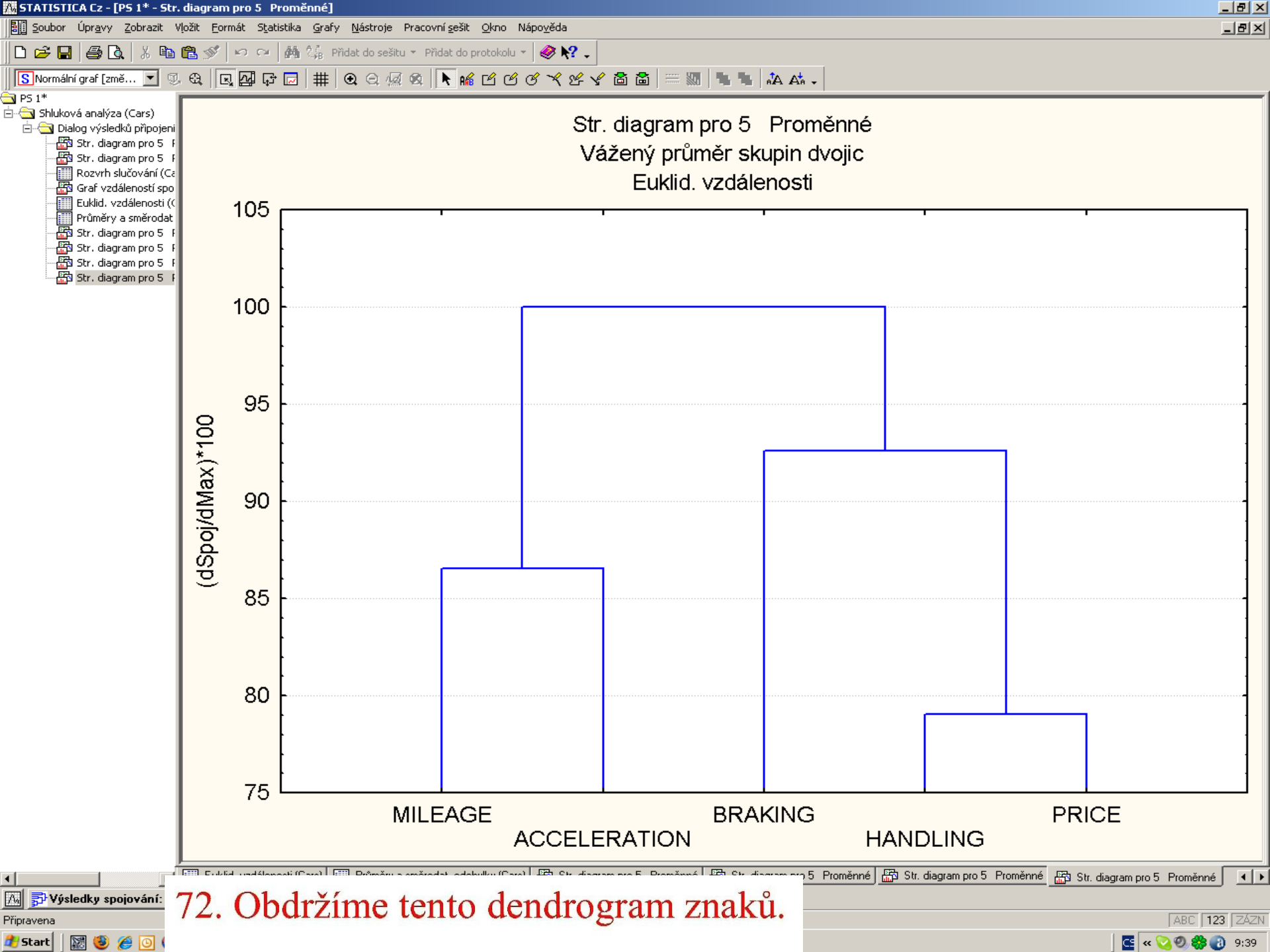


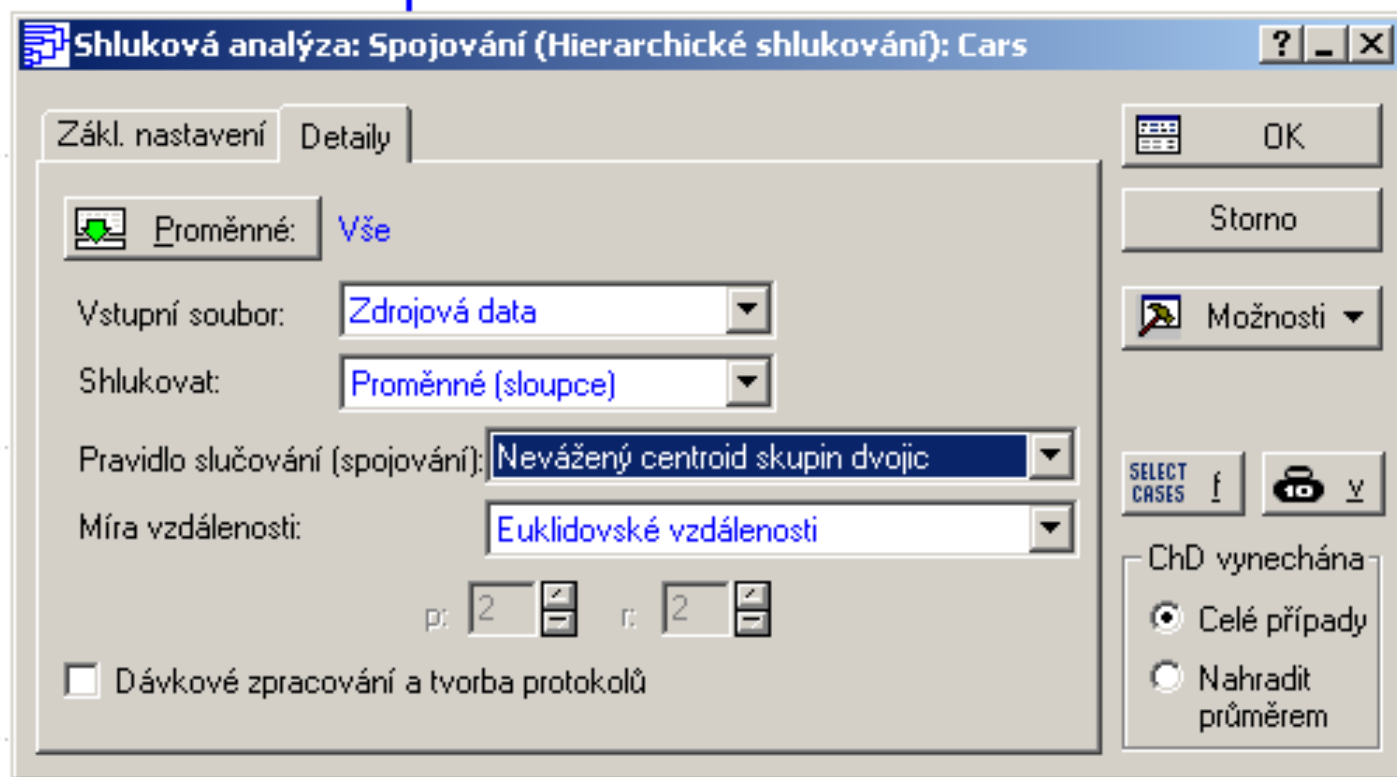
69. Navrátíme se a zadáme jinou shlukovací metodu *Nevážený průměr skupin dvojic* a OK.



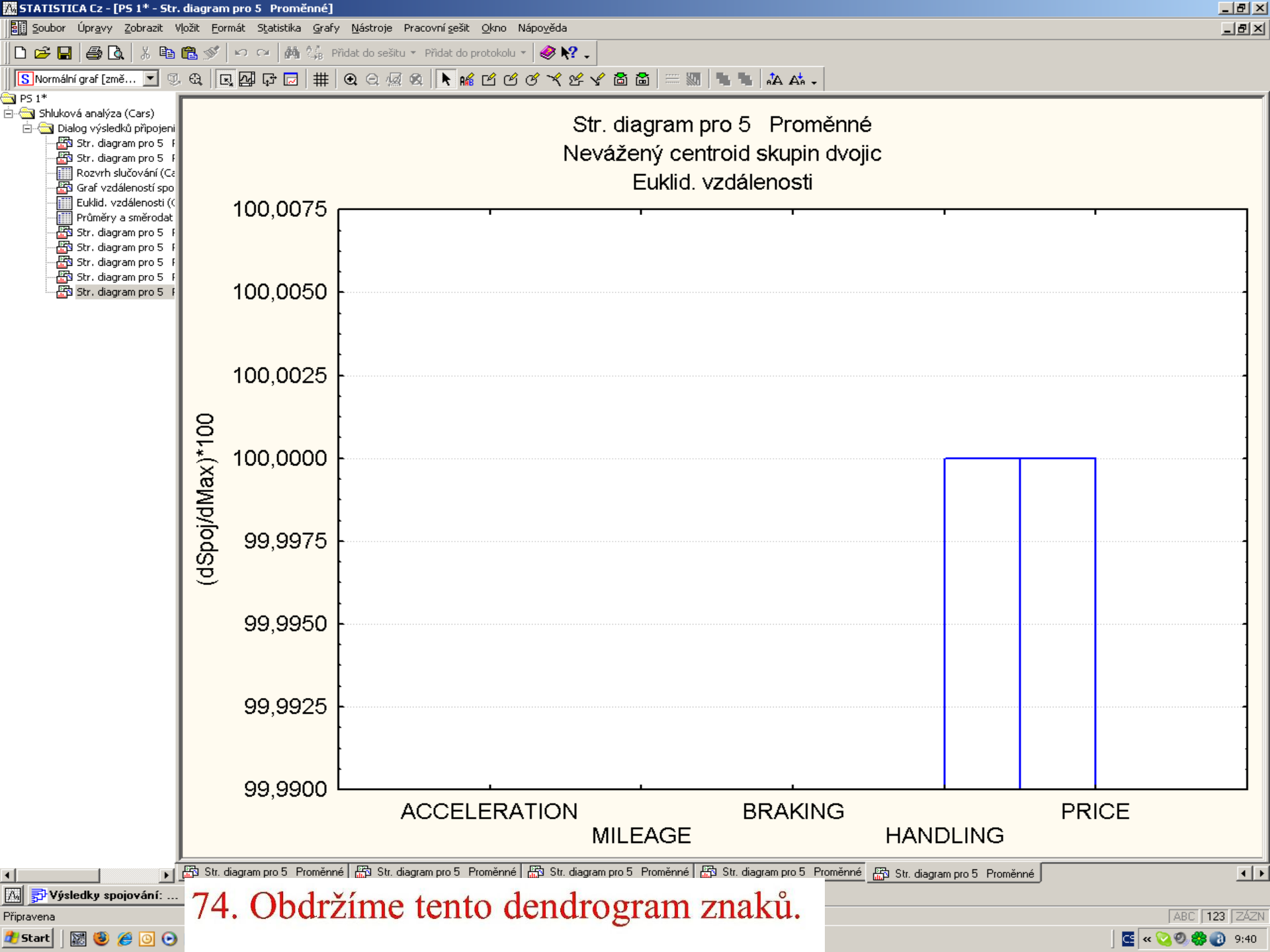


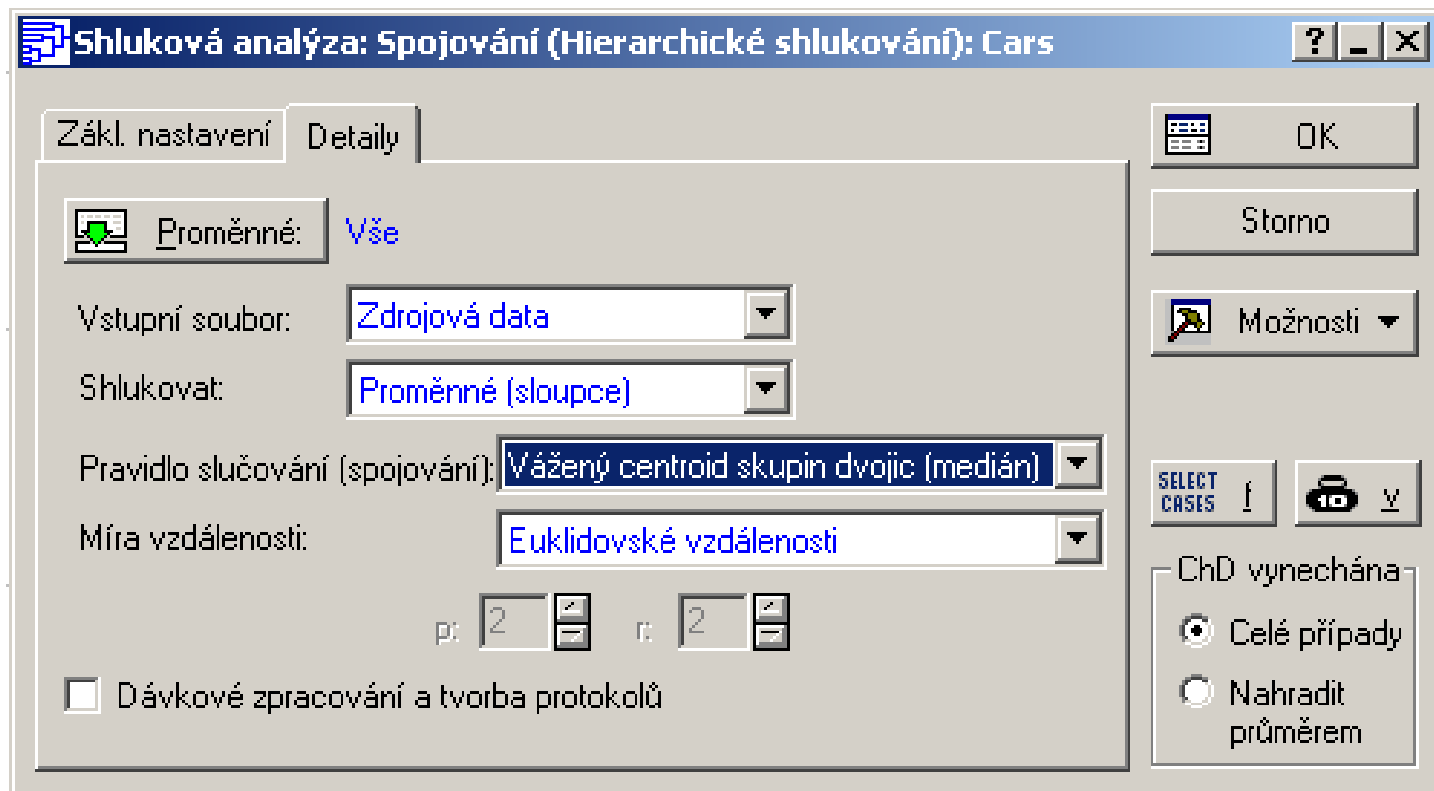
71. Navrátíme se a zadáme jinou shlukovací metodu *Vážený průměr skupin dvojic* a OK.



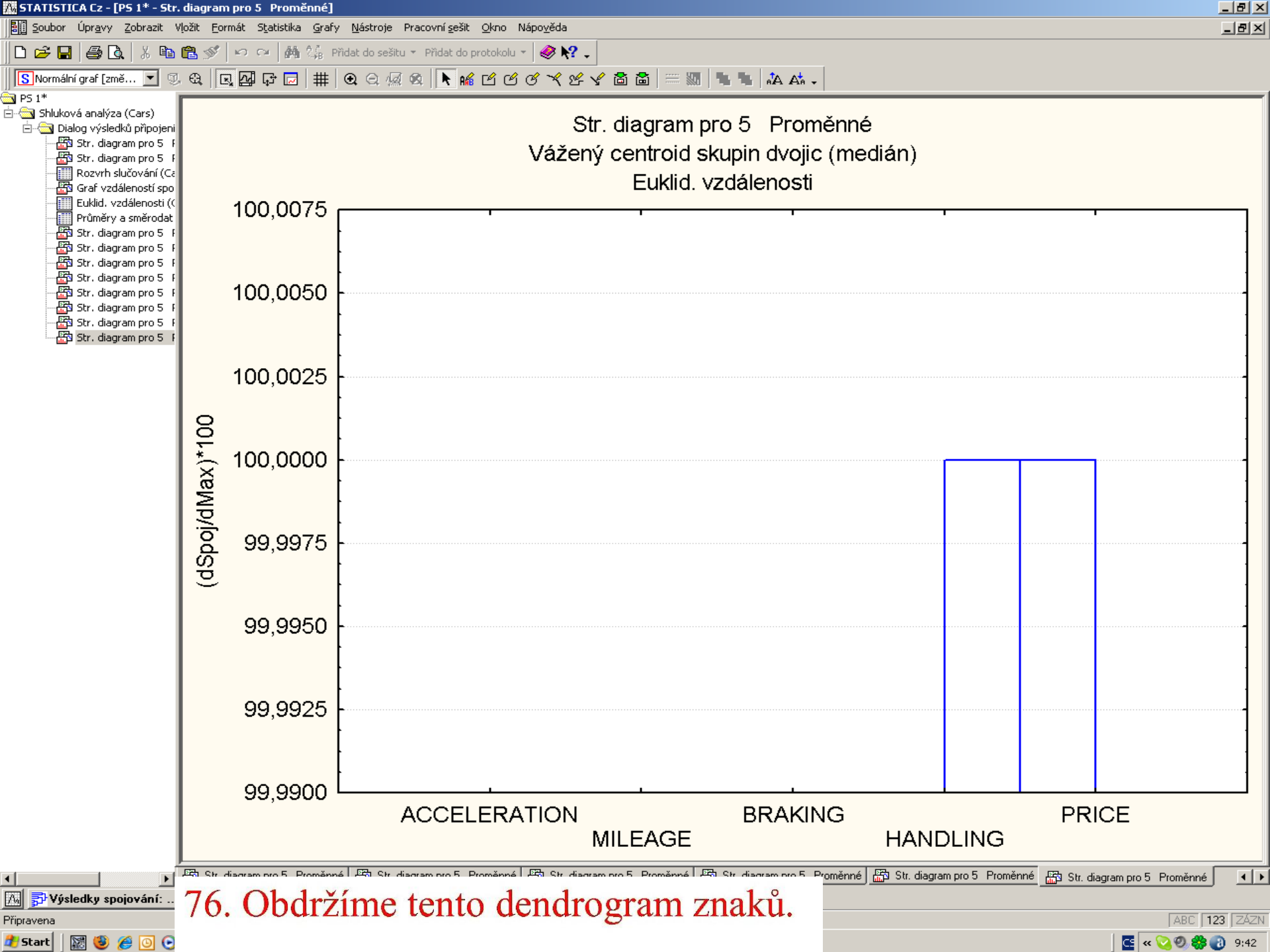


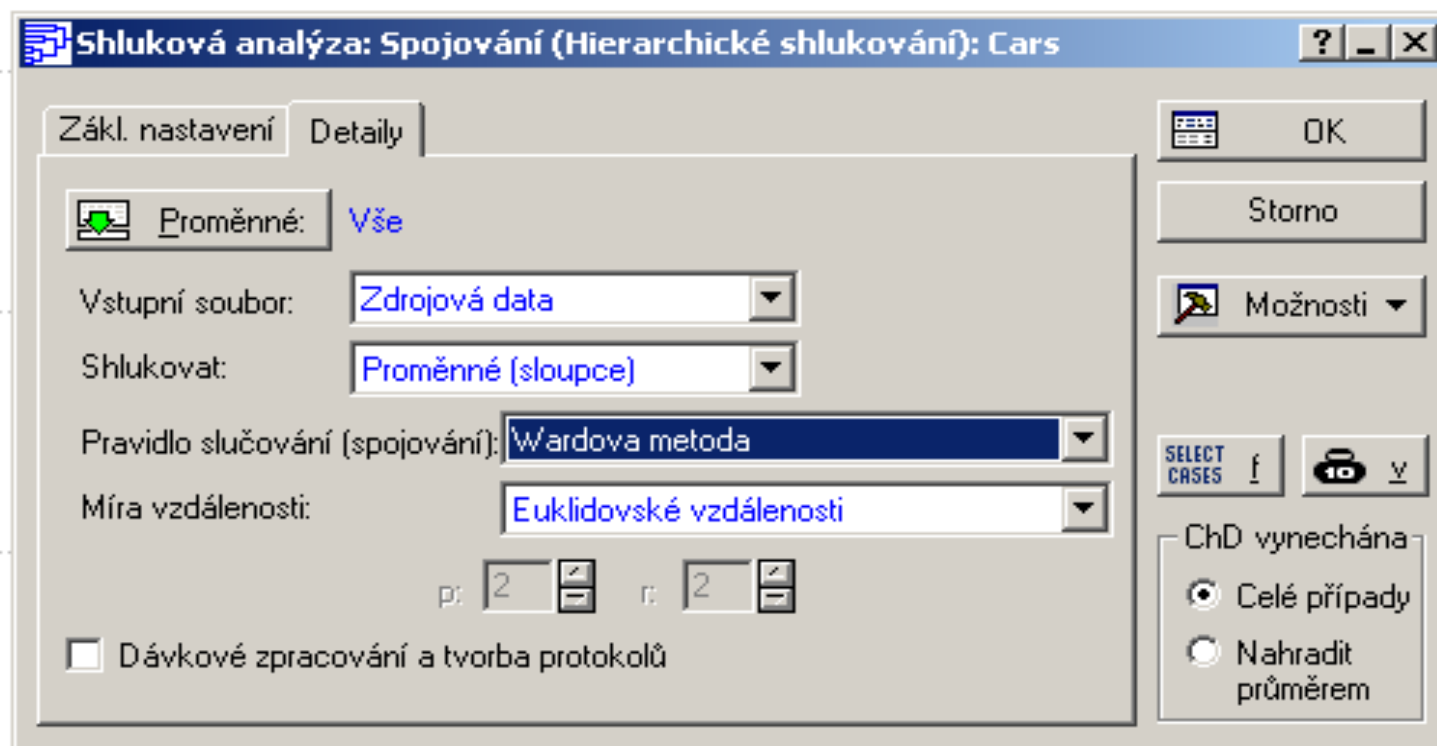
73. Navrátíme se a zadáme jinou shlukovací metodu *Nevážený centroid skupin dvojic* a OK.



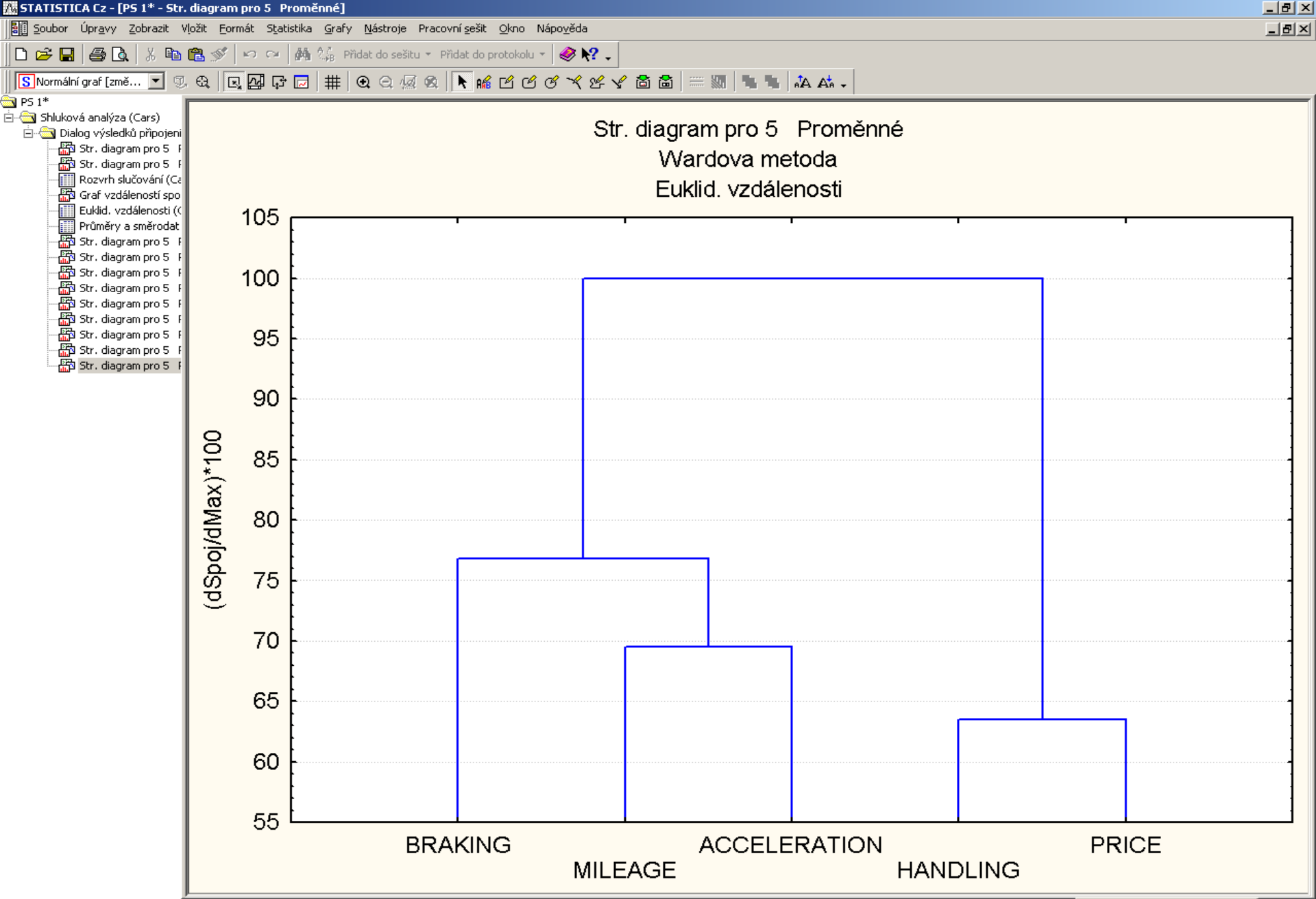


75. Navrátíme se a zadáme jinou shlukovací metodu *Vážený centroid skupin dvojic (medián)* a OK.





77. Navrátíme se a zadáme jinou shlukovací metodu *Wardovu* a OK.



78. Obdržíme tento dendrogram znaků.